Aalto-yliopiston perustieteiden korkeakoulu
Matematiikan ja systeemianalyysin laitos

Matias Leppisaari

Katastrofi- ja finanssiriskin mittaamisesta ääriarvoteoriaa soveltaen

Lisensiaatintutkimus

Helsinki, 11.6.2013

Vastuuprofessori: Prof. Ahti Salo

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of licentiate thesis

Author Matias Leppisaari

Title of thesis Measuring Catastrophe Risk and Financial Risk using Extreme Value Theory

Department Department of Mathematics and Systems Analysis

Field of research Systems and Operations Research

Supervising professor Prof. Ahti Salo **Code of professorship** F017Z

Thesis advisor(s) Prof. Ahti Salo

Thesis examiner(s) Prof. Juho Kanniainen

Number of pages 223 **Language** Finnish

Date of submission for examination 11.06.2013

Abstract

This Thesis considers Extreme Value Theory (EVT) and its statistical applications. Its aim is twofold: first, to give an introduction to statistical methods based on EVT; and second, to apply these methods to real-world modelling challenges. The presented applications are relevant both from the specific application area's point of view, as well as from a wider risk management perspective. In fact, one unifying theme in all the applications is their relevance to a (life, non-life, and/or re-) insurance company's management of insurance and financial risks.

The three applications considered in this Thesis are as follows. First, we examine the modelling of sea-level maxima using data provided by the Finnish Meteorological Institute, consisting of daily maximum sea-levels between 1904–2011 recorded at the coast of Helsinki. The sea-level modelling exercise is also used as a unified framework to illustrate the statistical methods of EVT considered in the thesis. The second application is to catastrophe modelling by using EVT to model the probability distribution of death counts caused by accidents or catastrophes in the Finnish population. The third application is concerned with realistic modelling of market risk and estimation of conditional risk measures (Value-at-Risk, Expected Shortfall).

We discuss risk measurement based on the estimated models, and compare the results from the accidental death model and equity risk model to the corresponding specifications used in the Quantitative Impact Study 5 (QIS 5) and Long Term Guarantees Assessment (LTGA) of the European Commission's Solvency II insurance directive. According to our catastrophe death model, the risk level implied by the mass accident scenario of LTGA seems to be significantly higher than the target 1-year 99.5 % Value-at-Risk (VaR) specified by Solvency II — at least for Finland. The risk level implied by the "arena risk" scenario of QIS 5 is higher still.

The models for extreme sea-level values of Baltic Sea clearly indicate that there has been an increasing trend in the sea-level *maxima* since 1904, even after accounting for the changes in the mean sea-level. The statistical models for sea-levels are used to provide estimates about the minimum required construction height at the coast of Helsinki, defined as the sea-level height that is exceeded on average once in the next 200 years.

This study does not contain new methodological contributions; instead, the focus is on applications. To the author's knowledge, sea-level maxima of the Baltic Sea have not been previously modelled using non-homogenous Poisson point processes and explanatory meteorological variables (in this case, the North Atlantic Oscillation or NAO-index). The application of EVT to the modelling of Finnish catastrophic accidental deaths is also new, whereas the market risk application is based on existing literature.

Keywords extreme value theory, EVT, GEV, GPD, generalized Pareto distribution, point processes, statistical modelling, extremes of random phenomena, catastrophe risk, market risk, risk measures, Value-at-Risk, sea-level maxima, accidental deaths, insurance risk, Solvency II

Aalto-yliopisto
Perustieteiden
korkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi

Lisensiaatintutkimuksen tiivistelmä

Tekijä Matias Leppisaari

Työn nimi Katastrofi- ja finanssiriskin mittaamisesta ääriarvoteoriaa soveltaen

Laitos Matematiikan ja systeemianalyysin laitos

Tutkimusala Systeemi- ja operaatiotutkimus

Vastuuprofessori Prof. Ahti Salo **Professuurikoodi** F017Z

Työn ohjaajat Prof. Ahti Salo

Työn tarkastajat Prof. Juho Kanniainen

Jätetty tarkastettavaksi 11.06.2013 **Sivumäärä** 223 **Kieli** suomi

Tiivistelmä

Tämä tutkimus käsittelee ääriarvoteoriaa (extreme value theory, EVT) ja sen tilastollisia sovelluksia. Työssä luodaan ensinnäkin katsaus ääriarvoteoriaan perustuviin tilastollisiin menetelmiin, ja toiseksi sovelletaan näitä menetelmiä reaalimaailman mallinnushaasteisiin. Esitetyt sovellukset ovat relevantteja paitsi kunkin sovellusalueen näkökulmasta, myös laajemmin riskienhallinnan perspektiivistä. Yhtenä yhdistävänä tekijänä kaikissa sovelluksissa onkin niiden relevanssi (henki-, vahinko- ja/tai jälleen-) vakuutusyhtiön vakuutus- ja finanssiriskien hallinnan kannalta.

Työssä tarkastellaan kolmea sovellusta. Ensimmäinen näistä käsittelee meren pinnankorkeuden maksimiarvojen mallintamista hyödyntäen Ilmatieteen laitokselta saatua, aikavälin 1904–2011 kattavaa päiväkohtaista mittausdataa vedenkorkeuksista Helsingin edustalla. Vedenkorkeusmaksimien mallinnusta käytetään myös havainnollistamaan työssä käsiteltyjä tilastollisia EVT-menetelmiä yhtenäisessä viitekehyksessä. Toisessa sovelluksessa käsitellään katastrofiriskin mallinnusta, ja käytetään ääriarvoteoriaa Suomen väestöä kohdanneiden onnettomuus- tai katastrofikuolemien lukumäärien todennäköisyysjakauman estimointiin. Kolmas sovellus koskee markkinariskin realistista mallinnusta ja ehdollisten riskimittojen (Value-at-Risk, Expected Shortfall) estimointia.

Tutkimuksessa käsitellään riskin mittaamista estimoituihin malleihin perustuen, ja verrataan tuloksia onnettomuuskuolemamallista sekä osakeriskimallista vastavaaviin, Euroopan komission Solvenssi II –vakuutusdirektiivin viidennessä vaikuttavuusarvioinnissa (QIS 5) ja LTGA-arviointipaketissa käytettyihin spesifikaatioihin näille riskeille. Erityisesti havaitaan, että tässä tutkimuksessa esitetyn katastrofikuolemamallin perusteella LTGA-vaikuttavuusarvion joukko-onnettomuusskenaarion riskitaso vaikuttaa selvästi Solvenssi II:ssa tavoitteena olevaa 1-vuoden 99.5 % Value-at-Risk (VaR) –mittaa korkeammalta – ainakin Suomen kohdalla. QIS5:n ”areenariskiskenaarion” riskitaso on vielä edellistäkin suurempi.

Äärimmäisille Itämeren vedenkorkeuden arvoille rakennetut mallit viittaavat vahvasti kasvavaan trendiin merenpinnan korkeuden *maksimiarvoissa*, myös merenpinnan keskimääräisen tason muutoksien huomioimisen jälkeen. Malleja käytetään arvioimaan Helsingin rannikolla edellytettyä pienintä sallittua rakennuskorkeutta, joka määritellään sellaiseksi vedenkorkeuden arvoksi, joka ylitetään keskimäärin kerran seuraavan 200 vuoden aikana.

Tämä tutkimus ei sisällä uusia menetelmäkehitykseen liittyviä kontribuutioita, vaan esityksen painotus on sovelluksissa. Kirjoittajan tietämän mukaan Itämeren vedenkorkeuden maksimeja ei ole aiemmin mallinnettu tutkimuksen luvun 2 tapaan epähomogeenisia pisteprosesseja ja meteorologisia selittäviä muuttujia (tässä tapauksessa NAO-indeksiä) käyttäen. Onnettomuuskuolemamäärien mallintaminen ääriarvoteoriaa käyttäen eli luvun 3 sovellus on myös uusi. Luvussa 4 esitetty lähestymistapa markkinariskin mallinnukseen puolestaan noudattaa kirjallisuudessa aiemmin esitettyä.

Avainsanat ääriarvoteoria, EVT, GEV, GPD, yleistetty Pareto-jakauma, pisteprosessit, tilastollinen mallinnus, satunnaisilmiöiden ääriarvot, katastrofiriski, markkinariski, riskimitat, Value-at-Risk, vedenkorkeusmaksimit, onnettomuuskuolemat, vakuutusriski, Solvenssi II

Alkusanat

Tämä tutkimus pohjautuu allekirjoittaneen aiemmin Suomen Aktuaariyhdistyksen Working Papers -sarjassa julkaistuun käsikirjoitukseen (SHV-työ). Haluan kiittää FM, SHV Tapani Tuomista mielenkiintoisesta keskustelusta ja rakentavista kommenteista koskien erityisesti tutkimuksen toista sovellusta, katastrofikuolemien mallinnusta.

Erityisesti haluan käyttää hyväkseni tilaisuuden ja kiittää tässä kirjallisesti vaimoani Hanaa, joka on vankkumattomasti tukenut minua kaikessa – huolimatta lukuisista illoista, jotka päädyin viettämään *kiehtovan* kirjallisuuden ja *hauskan* MATLAB-koodauksen parissa, pyrkiessäni viemään jatko- ja muita opintoja eteenpäin työn ohessa.

Helsingissä 11.6.2013,

Matias Leppisaari

Sisältö

Kuvat

Taulukot

Johdanto

Ääriarvoteoria (Extreme Value Theory, EVT) tarkastelee satunnaisilmiöiden ääriarvoja. Sitä voitaisiin kutsua myös äärimmäisten ilmiöiden teoriaksi, missä "äärimmäisellä" ymmärretään matemaattisessa mielessä tietyn ilmiön *maksimi*-arvojen ja niiden jakauman, tai taustalla olevan todennäköisyysjakauman *häntien*, tarkastelemista.

Ääriarvoteorian piiriin kuuluvien menetelmien ja niiden erikoistapauksien soveltamisella tilastolliseen dataan on pitkät perinteet hydrologiassa (vesistötieteessä), luotettavuustekniikassa (reliability engineering) ja vakuutusmatematiikassa, vaikka menetelmiä ei aina tällä nimellä olekaan kutsuttu. Parin viime vuosikymmenen aikana sovelluksien alue on kuitenkin laajentunut valtavasti, käsittäen esimerkiksi telekommunikaation, ilmastotieteen, aerodynamiikan, lääketieteellisen tekniikan, seismologian, ilmakehän kemian, tähtifysiikan ja materiaalitekniikan; ks. esim. [1, 2]. Listaa voisi jatkaa loputtomiin. Erään tärkeimmistä uusista sovellusalueista menetelmien kehittymisen (eikä pelkästään olemassaolevien soveltamisen) kannalta muodostavat finanssimarkkinat ja finanssiriskienhallinta. Finanssi- ja vakuutusala lieneekin nykyään ääriarvoteorian soveltavan tutkimuksen merkittävimpiä painopistealueita, mahdollisesti heti ympäristötieteiden (kaikissa eri muodoissaan) jälkeen.

Riskiä mallinnetaan matemaattisesti pohjimmiltaan valitsemalla tietty *todennäköisyysjakauma* kuvailtavalle ilmiölle.[1] Usein todennäköisyysjakauma on estimoitu havaintojen perusteella tilastollista analyysiä käyttäen. Useimmissa tilastollisissa menetelmissä kiinnostus kohdistuu pääasiallisesti tarkasteltavien havaintojen taustalla olevan todennäköisyysjakauman keskiosiin. Jakauman häntiin eli äärimmäisiin arvoihin joko havaintojen ylä- tai alapäässä ei sen sijaan useinkaan kiinnitetä menetelmissä erityistä huomiota, tai tällaisia arvoja pidetään jopa poikkeavina havaintoina (outlier). On kuitenkin tilanteita, joissa äärim-mäiset havaintoarvot ovat tärkein osa probleemaa.

Ääriarvoteoriaa sovellettaessa kiinnostus kohdistuu nimenomaan ilmiön tai prosessin äärimmäisiin arvoihin. Yleensä ääriarvoanalyysin tavoitteena on arvioida kaikkia tähän mennessä havaittuja tapahtumia suurempien (tai pienempien) tapahtumien todennäköisyyksiä – eli ekstrapoloida havaintojen ulkopuolelle. Ääriarvoteoria voidaan nähdä työkaluna, joka tarjoaa keinot ilmiön taustalla olevan todennäköisyysjakauman häntätodennäköisyyksien estimoimiseen niin

[1] Tämä jakauma voi olla esimerkiksi ajassa muuttuva, eikä aina ole ilmaistavissa analyyttisessä muodossa. Usein se on myös monien mallinnuksessa tehtyjen osavalintojen implisiittisesti määräämä.

hyvin kuin havaintojen perusteella on mahdollista. Toisaalta, vaikkei saatavilla olisi ollenkaan käyttökelpoista dataa, teorian perusteella voidaan myös osoittaa, minkä tyyppisiä jakaumia tulisi käyttää, jotta katastrofiriskien mahdollisuus tulee huomioitua konservatiivisesti. Ääriarvoteorian soveltuvuutta käytännön ongelmiin kuvaa hyvin seuraava, teoksesta [2] otettu Richard Smithin toteamus: *"There is always going to be an element of doubt, as one is extrapolating into areas one doesn't know about. But what EVT is doing is making the best use of whatever data you have about extreme phenomena."*

Tämän esityksen tavoite on kahtalainen: ensinnäkin, luoda katsaus ääriarvoteorian tilastollisiin sovelluksiin, mukaan lukien lyhyt katsaus menetelmien pohjana olevaan matemaattiseen teoriaan, ja toiseksi, esittää konkreettisia reaalimaailman sovelluksia jotka havainnollistavat teorian tarjoamia menetelmiä myös käytännössä. Esitetyt sovellukset ovat relevantteja paitsi kunkin sovellusalueen näkökulmasta, myös laajemmin riskienhallinnan perspektiivistä. Yhtenä yhdistävänä tekijänä kaikissa sovelluksissa onkin niiden relevanssi (henki-, vahinko- ja/tai jälleen-) vakuutusyhtiön vakuutus- ja finanssiriskien hallinnan kannalta.

Tilastollisia menetelmiä käytettäessä on aina kiinnitettävä huomiota menetelmien taustalla oleviin oletuksiin sekä menetelmien soveltuvuusalueeseen ja rajoituksiin. Aivan erityisesti tämä pätee ääriarvoteoriaan perustuvien tilastollisten menetelmien käytössä, kun pyritään kuvaamaan äärimmäisiä ilmiöitä, jotka jo määritelmän mukaan ovat harvinaisia ja joista on usein vain vähän tai ei ollenkaan havaintoja olemassa. Esityksessä tullaan menetelmien soveltamisen yhteydessä kiinnittämään huomiota taustaoletuksiin ja niiden toteutumiseen, sekä painottamaan graafisen tarkastelun hyödyllisyyttä (ja välttämättömyyttä) reaalimaailman dataa analysoitaessa.

Esityksen rakenne on seuraava. Luvussa 1 luodaan lyhyt katsaus matemaattiseen taustateoriaan, siltä osin kuin myöhemmin esitettävien menetelmien ymmärtämiseksi tai perustelemiseksi on tarpeen. Esityksen painopisteenä olevat tilastolliset menetelmät esitetään luvussa 2 likimain historiallisessa (ja myös vaikeus-) järjestyksessä. Luvussa käytetään esimerkkinä Itämeren vedenkorkeuden mallinnusta Helsingin edustalla, ja vedenkorkeusilmiötä analysoidaan yksityiskohtaisesti esitettyjä menetelmiä soveltamalla. Seuraavissa luvuissa esitetään kaksi muuta sovellusta: Luvussa 3 tarkastellaan onnettomuuskuolemien määrien mallintamista, ml. onnettomuuskuolemien simulointia, ja luvussa 4 puolestaan realistista markkinariskin mallintamista ja riskimittojen (Value-at-Risk, Expected Shortfall) estimointia. Onnettomuuskuolemamallin ja osakeriskimallin antamia tuloksia verrataan myös Solvenssi II:n viidennessä vaikuttavuusarvioinnissa (QIS 5) ja uudemmassa LTGA-arviointipaketissa käytettyihin vastaaviin spesifikaatioihin osakemarkkinariskin ja sairaus- ja tapaturmavakuutuksen katastrofiriskin (konsentraatioriski ja "areena"- tai joukko-onnettomuusriski) osalta. Luku 5 sisältää yhteenvedon.

Luku 1

Taustateoriaa

Tässä luvussa luodaan katsaus ääriarvoteoriaan, painottaen erityisesti sovellusten kannalta tärkeitä tuloksia. Tavoitteena on antaa lukijalle käsitys teorian perusteista, tai teoriaan perehtyneen lukijan osalta toimia aiheen kertauksena.

Esitys perustuu pääasiassa lähteisiin [2] ja [5]. Tulokset esitetään ilman todistuksia, todistusten osalta viitataan lähdekirjallisuuteen (Embrechts et. al [2], Resnick [3], Leadbetter et al. [4]).

Osiossa 1.1 luodaan ensin yhteenvedonomainen katsaus joihinkin mittateorian ja todennäköisyysteorian käsitteisiin. Osio koostuu lähinnä käsitteiden määrittelyistä ja on tarkoitettu taustamateriaaliksi; osion voi huoletta ohittaa, ja palata siihen tarvittaessa. Johdanto varsinaiseen ääriarvoteoriaan alkaa osiosta 1.2, ja kohdassa 1.3 esitetään ääriarvojakaumat eli niiden jakaumien luokka, johon satunnaismuuttujien maksimit suppenevat jakauman suhteen. Osiossa 1.4 käsitellään ääriarvoteoriaa stationaarisille prosesseille, ja osiossa 1.5 yleistettyä Pareto-jakaumaa mallina kaikkien tietyn korkean tason ylitteiden jakaumalle. Osiossa 1.6 esitetään pisteprosessit mallina ääriarvojen kuvaamiseen. Pisteprosessilähestymistapa mahdollistaa myös epästationaaristen ilmiöiden ja prosessien mallintamisen luonnollisella tavalla. Luvun lopussa kohta 1.7 sisältää hyvin lyhyen katsauksen ääriarvoteorian historiaan (lähinnä sovelluksien näkökulmasta), ja osiossa on esitetty myös historiallinen esimerkki motivaatioksi luvun 2 tilastollisiin sovelluksiin.

1.1 Yleistä mitta- ja todennäköisyysteoriasta

Aloitetaan esittämällä lyhyesti joitakin käsitteitä mittateoriasta. Tässä esityksessä mittateoriaa tarvitaan lähinnä pisteprosesseja (osio 1.6 ja liite D) koskevien tulosten perustelemiseksi, mutta esitys on sijoitettu tähän yhteyteen, koska se mahdollistaa todennäköisyyteen liittyvien käsitteiden määrittelyn täsmällisesti alaosiossa 1.1.2.

Todennäköisyysteorian osalta varsin helposti lähestyttävä perusteos on Jacod &

Protter [6]. Billingsley [7] on klassikko, ja Durrett [8] uudempi erinomainen esitys todennäköisyysteoriasta. Standardeja kirjallisuusviitteitä stokastisten prosessien osalta ovat Karatzas & Shreve [9] ja Revuz & Yor [10] (tässä esityksessä ei kuitenkaan tarvita martingaaleja, stokastisia differentiaaliyhtälöjä tai stokastista integrointia). Mittateorian ja analyysiin liittyvien tulosten osalta viitataan teokseen Gariepy & Ziemer [11].

Seuraava esitys noudattelee lähteen [12] esitystapaa.

1.1.1 Mittateoriaa

Avaruus X täysin yleisesti on mielivaltainen joukko. Sanaa "avaruus" käytetään usein osoittamaan joukkoa, joka on varustettu erityisellä rakenteella. Esimerkiksi vektoriavaruus on joukko, esim. $\mathbb{R}^n$, joka on varustettu algebrallisella rakenteella. Termejä "perhe" ja "kokoelma" käytetään yleisesti joukon synonyymeinä.

Topologinen avaruus ja metrinen avaruus. Topologiset avaruudet ovat joukkoja, joiden rakenne mahdollistaa puhumisen sellaisista käsitteistä kuin suppeneminen ja jatkuvuus. Metrinen avaruus on lisäksi varustettu metriikalla, joka mahdollistaa "etäisyyksistä" puhumisen kvantitatiivisesti.

Määritelmä 1.1 (Topologinen avaruus)
Paria $(X, \mathscr{T})$*, missä* X *on ei-tyhjä joukko ja* $\mathscr{T}$ *on* X*:n osajoukkojen perhe, kutsutaan* topologiseksi avaruudeksi*, kun seuraavat ehdot täyttyvät:*

(i) Tyhjä joukko $\emptyset$ *ja koko avaruus* X *kuuluvat* $\mathscr{T}$*:hen:* $\emptyset, X \in \mathscr{T}$*.*

(ii) Jos $\mathcal{S}$ *on mielivaltainen* $\mathscr{T}$*:n alikokoelma, niin*

$$\bigcup \{U : U \in \mathcal{S}\} \in \mathscr{T}.$$

(iii) Jos $\mathcal{S}$ *on mikä tahansa* $\mathscr{T}$*:n äärellinen alikokoelma, niin*

$$\bigcap \{U : U \in \mathcal{S}\} \in \mathscr{T}.$$

Kokoelmaa $\mathscr{T}$ kutsutaan avaruuden X *topologiaksi* ja $\mathscr{T}$:n elementit ovat X:n avoimia joukkoja. Avointa joukkoa, joka sisältää pisteen x, kutsutaan x:n ympäristöksi. Joukon $A \subset X$ sisus $\mathrm{int}(A)$ tai A° on kaikkien A:han sisältyvien avoimien joukkojen unioni; $\mathrm{int}(A)$ on avoin joukko, ja on myös mahdollista, että jonkin joukon sisus on tyhjä joukko. Joukkoa A kutsutaan suljetuksi, jos sen komplementti $A^c = X \backslash A$ on avoin. Joukon A sulkeuma on $\mathrm{cl}(A) = \overline{A} = X \backslash \mathrm{int}(X \backslash A)$, ja sen reuna on $\partial A = \overline{A} \backslash A$. Sulkeumalle pätee $A \subset \overline{A}$.

Olkoon $(X, \mathscr{T})$ topologinen avaruus. Avaruuden X *kanta* B on sellainen avoimien joukkojen kokoelma $\mathscr{T}$:ssä, että jokainen $\mathscr{T}$:n avoin joukko voidaan kirjoittaa B:n elementtien unionina. Jos jokainen avoin joukko voidaan kirjoittaa B:n elementtien numeroituvana unionina, on B avaruuden numeroituva kanta. Sanotaan, että kanta B *generoi* topologian $\mathscr{T}$. Kannan käsite on erityisen hyödyllinen, koska monet topologiat on helpointa määritellä topologiat generoivien kantojen kautta, ja toisaalta monet topologian ominaisuudet voidaan palauttaa topologian generoivan kannan ominaisuuksiin.

Jatketaan määritelmillä.

Määritelmä 1.2 (Hausdorff-avaruus)
Topologisen avaruuden X sanotaan olevan Hausdorff-avaruus *(*Hausdorff*) jos jokaiselle erillisten pisteiden parille $x_1, x_2 \in X$ on olemassa erilliset avoimet joukot U_1 ja U_2 (ts. $U_1 \cap U_2 = \emptyset$) s.e. $x_1 \in U_1$ ja $x_2 \in U_2$.*

Toisin sanoen avaruus on Hausdorff, jos sen mitkä tahansa kaksi erillista pistettä voidaan *erottaa* erillisillä avoimilla joukoilla.

Seuraavaa määritelmää tarvitaan kompaktin joukon määrittelemiseksi.

Määritelmä 1.3 (Avoin peite ja osapeite)
Olkoon $(X, \mathscr{T})$ topologinen avaruus. Avoimien joukkojen kokoelma $\mathcal{U} = \{U_\alpha | \alpha \in I\}$ on joukon $A \subset X$ avoin peite*, jos*

$$A \subset \bigcup \mathcal{U} = \bigcup_{\alpha \in I} U_\alpha.$$

Tällöin sanotaan lisäksi, että $\mathcal{S} \subset \mathcal{U}$ on joukon A osapeite*, jos myös $A \subset \bigcup \mathcal{S}$.*

Määritelmä 1.4 (Kompakti joukko)
Joukko $K \subset X$ on kompakti*, jos sen jokaisella avoimella peitteellä on olemassa äärellinen osapeite.*

Avaruuden X sanotaan olevan lokaalisti kompakti, jos jokainen X:n piste x sisältyy johonkin avoimeen joukkoon, jonka sulkeuma on kompakti.

Topologisen avaruuden X joukko $Y \subset X$ on suhteellisesti kompakti (relatively compact) joukko tai aliavaruus, jos Y:n sulkeuma on kompakti. Koska kompaktin avaruuden suljetut joukot ovat kompakteja joukkoja, jokainen kompaktin avaruuden osajoukko on suhteellisesti kompakti.

Topologisen avaruuden ohella toinen tärkeä käsite on metrinen avaruus, ja siihen liittyvä metriikka.

Määritelmä 1.5 (Metrinen avaruus)
Metrinen avaruus (X, d) *on mielivaltainen joukko X varustettuna* metriikalla *$d : X \times X \to [0, \infty)$, joka täyttää seuraavat ehdot kaikilla $x, y, z \in X$:*

(i) $d(x, y) = 0 \iff x = y$,

(ii) $d(x, y) = d(y, x)$,

(iii) $d(x, y) \leq d(x, z) + d(z, y)$.

Metriikka mahdollistaa etäisyyskäsitteen kvantitatiivisen määrittelyn, ja $d(x, y)$ on "etäisyys x:stä y:hyn". Ehtoa (iii) kutsutaan kolmioepäyhtälöksi.

Esimerkki 1.6 (Euklidinen metriikka)
Olkoon $X = \mathbb{R}^n$ ja $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Asetetaan $d : \mathbb{R}^n \times \mathbb{R}^n \to [0, \infty)$,

$$d(x, y) = \|x - y\| := \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2}.$$

Metrisessä avaruudessa (X, d) pätee seuraava tulos: jos joukko $K \subset X$ on kompakti, K on rajoitettu ja suljettu. Metrisessä topologiassa siis kompakti $\Rightarrow$ suljettu & rajoitettu, mutta implikaatio ei päde toiseen suuntaan. Sen sijaan euklidisessa topologiassa kompakti $\iff$ suljettu & rajoitettu.

Sigma-algebrat ja mitallisuus. Esitetään seuraavaksi keskeiset sigma-algebran ja mitallisuuden käsitteet. Olkoon $\mathcal{P}(X) = \{A \mid A \subset X\}$ joukon X kaikkien osajoukkojen kokoelma eli X:n potenssijoukko.

Määritelmä 1.7 (Sigma-algebra)
*Kokoelma $\mathcal{M}$ on (joukon X) σ-*algebra*, jos seuraavat ehdot täyttyvät:*

(i) $\emptyset \in \mathcal{M}$ *ja* $X \in \mathcal{M}$,

(ii) $A \in \mathcal{M} \iff A^c \in \mathcal{M}$,

(iii) $\{A_j\}_{j=1}^{\infty} \subset \mathcal{M} \Rightarrow \bigcup_{j=1}^{\infty} A_j \in \mathcal{M}$.

σ-algebrat ovat siis komplementin ottamisen ja numeroituvien unioneiden suhteen suljettujen joukkojen luokka. Rajoitus numeroituviin unioneihin kohdassa (iii) on olennainen, ylinumeroituvia unioneita ei sallita tässä. Määritelmästä seuraa suoraan, että σ-algebra on suljettu myös äärellisten leikkausten suhteen, sillä $\cap_{j=1}^{\infty} A_j = \left(\cup_{j=1}^{\infty} A_j^c\right)^c$.

Paria $(X, \mathcal{M})$, missä X on avaruus ja $\mathcal{M}$ σ-algebra X:llä, kutsutaan *mitalliseksi avaruudeksi*. Sigma-algebran $\mathcal{M}$ elementit ovat *mitallisia joukkoja*. Olkoon $(X, \mathcal{M})$, $(Y, \mathcal{N})$ kaksi mitallista avaruutta ja $f : X \to Y$ kuvaus eli funktio avaruudesta X avaruuteen Y. Kuvaus f on *mitallinen*, jos kaikilla $B \in \mathcal{N}$ joukon B alkukuva

$$f^{-1}(B) = \{x \in X \mid f(x) \in B\} = \{f \in B\}$$

kuuluu $\mathcal{M}$:ään, $f^{-1}(B) \in \mathcal{M}$. Olkoon $g : Y \to Z$ toinen kuvaus, jolloin $g \circ f : X \to Z$ on yhdistetty kuvaus ($\forall x \in X : (g \circ f)(x) := g(f(x))$). Mitallisuus säilyy funktioiden yhdisteissä $g \circ f$. Mitalliset funktiot (vast. joukot) ovat "hyviä" funktioita (joukkoja).

Sigma-algebrat ovat usein hyvin monimutkaisia, erityisesti ylinumeroituvissa avaruuksissa, eikä ole helppoa sääntöä sen testaamiseksi, kuuluuko tietty joukko σ-algebraan. Tavallinen tapa määritellä σ-algebra on ottaa pienempi kokoelma joukkoja, jotka voidaan eksplisiittisesti kuvata, ja generoida σ-algebra näistä eli varustaa avaruus näiden joukkojen generoimalla σ-algebralla.

Lemma 1.8 (σ-algebrojen leikkaus) *Olkoon $\{\mathcal{M}_j \mid j \in J\}$ kokoelma avaruuden X σ-algebroja. Tällöin leikkaus*

$$\mathcal{M} = \bigcap_{j \in J} \mathcal{M}_j$$

on myös σ-algebra.

Olkoon $\mathcal{A}$ mielivaltainen kokoelma X:n osajoukkoja. $\mathcal{A}$:n generoima sigma-algebra, merkitään $\sigma(\mathcal{A})$, on kaikkien niiden X:n σ-algebrojen leikkaus, jotka sisältävät $\mathcal{A}$:n. Määritelmästä seuraa, että mille tahansa X:n σ-algebralle $\mathcal{E}$ pätee $\mathcal{A} \subseteq \mathcal{E} \Rightarrow \sigma(\mathcal{A}) \subseteq \mathcal{E}$, joten $\sigma(\mathcal{A})$ on pienin σ-algebra X:llä, joka sisältää $\mathcal{A}$:n.

Vastaavasti voidaan puhua funktioiden generoimista σ-algebroista. Olkoon $(Y, \mathcal{N})$ mitallinen avaruus ja Φ kokoelma funktioita $X \to Y$. Tällöin kokoelman Φ generoima sigma-algebra määritellään

$$\sigma(\Phi) = \sigma\left\{f^{-1}(B) \mid f \in \Phi, B \in \mathcal{N}\right\},$$

ja se on pienin σ-algebra, joka tekee kaikista Φ:n funktioista mitallisia.

Määritelmä 1.9 (Borel-σ-algebrat) *Olkoon X metrinen avaruus.* Borel-σ-algebra $\mathscr{B}_X$ *on pienin σ-algebra X:llä, joka sisältää kaikki avoimet joukot. $\mathscr{B}_X$:n elementtejä kutsutaan* Borel-joukoiksi.

Vaihtoehtoisesti merkitään Borel-σ-algebraa $\mathscr{B}(X)$. Esimerkiksi reaaliakselin Borel-σ-algebran $\mathscr{B}_{\mathbb{R}}$ generoivat välit $\{(a, b] : -\infty < a < b < \infty\}$ tai välit $\{(-\infty, b) : -\infty < b < \infty\}$. Laajennetun reaaliakselin $\bar{\mathbb{R}}$ Borel-σ-algebran $\mathscr{B}_{\bar{\mathbb{R}}}$ generoivat vastaavasti muotoa $\{[-\infty, b] : b \in \mathbb{R}\}$ olevat välit.

Reaaliarvoisen funktion (tai satunnaismuuttujan, ks. jäljempänä) mitallisuudesta puhuttaessa avaruus varustetaan tässä esityksessä lähtökohtaisesti aina $\mathbb{R}$:n (tai $\bar{\mathbb{R}}$:n) Borel-sigma-algebralla.

Mitat. Olkoon $(X, \mathcal{M})$ mitallinen avaruus.

Määritelmä 1.10 (Mitta)
Kuvaus $\mu : \mathcal{M} \to [0, \infty]$ on mitta*, kun $\mathcal{M}$ on σ-algebra ja seuraavat ehdot pätevät:*

(i) $\mu(\emptyset) = 0$.

(ii) *Jos $\{A_j\}_{j=1}^{\infty} \subset \mathcal{M}$ on pistevieras (eli $A_j \cap A_k = \emptyset, j \neq k$), niin*

$$\mu\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mu(A_j).$$

Ominaisuus (ii) on ns. *numeroituva additiivisuus*. Kolmikkoa $(X, \mathcal{M}, \mu)$ kutsutaan *mitta-avaruudeksi*.

Jos $\mu(X) < \infty$, niin μ on *äärellinen mitta*. Mikäli $\mu(K) < \infty$ kaikilla kompakteilla joukoilla $K \subset X$, niin μ on *Radon-mitta*. Tällaiset mitat tulevat olemaan tärkeässä asemassa myöhemmin. Jos on olemassa mitalliset joukot $\{A_i\}$ s.e. $X = \bigcup A_i$ ja $\mu(A_i) < \infty \ \forall\, i$, mittaa μ kutsutaan σ*-äärelliseksi*. Mikäli $\mathcal{M}$ on Borel-σ-algebra metrisessä avaruudessa X, mittaa μ kutsutaan Borel-mitaksi. Jos mitalle pätee $\mu(X) = 1$, niin μ on todennäköisyysmitta, ks. seuraava alaosio.

Määritelmä 1.11 (Melkein kaikkialla)
Olkoon $(X, \mathcal{M}, \mu)$ mitta-avaruus. Sanotaan, että jokin ominaisuus pätee μ-melkein kaikkialla *(μ-m.k.), jos ominaisuus pätee joukossa $N^c = X \backslash N$, missä $N \in \mathcal{M}$ ja $\mu(N) = 0$.*

Ominaisuus pätee siis (μ-)melkein kaikkialla, kun se pätee kaikkialla paitsi (μ-)nollamittaisessa joukossa.

Mitta $\mu : \mathcal{M} \to [0, \infty]$ on *täydellinen*, jos kaikilla $B \in \mathcal{M}$ kaikilla $A \subset E$

$$(\mu(B) = 0 \ \wedge \ A \subset B) \Rightarrow A \in \mathcal{M}.$$

Siis mitta μ on täydellinen, kun kaikki (μ-)nollamittaisten joukkojen osajoukot ovat mitallisia, ja kolmikkoa $(X, \mathcal{M}, \mu)$ kutsutaan täydelliseksi mitta-avaruudeksi. Kaikki mitat eivät ole täydellisiä, mutta jokainen mitta voidaan suoraviivaisesti laajentaa täydelliseksi laajentamalla mitan määrittelyjoukko sisältämään kaikki nollamittaisten joukkojen osajoukot.

Esimerkki 1.12 (Lebesgue-Stieltjes–mitat) *Olkoon F ei-vähenevä, oikealta jatkuva reaaliarvoinen funktio $\mathbb{R}$:llä. Määritellään väleillä $(a, b]$, $a \leq b$, funktio*

$$\mu_0((a, b]) = F(b) - F(a).$$

Jos $a = -\infty$ tai $b = \infty$, asetetaan $F(\infty) := \lim_{x \nearrow \infty} F(x)$ ja $F(-\infty) := \lim_{y \searrow -\infty} F(y)$ (F voi saada arvon $\pm\infty$). Voidaan osoittaa, että on olemassa yksikäsitteinen mitta μ avaruudella $(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$ s.e. $\mu = \mu_0$, joka siis liittää "massan" $F(b) - F(a)$ kuhunkin väliin $(a, b]$. Tätä mittaa kutsutaan F:n Lebesgue-Stieltjes*–mitaksi $\mu = \mu_F$.*

Tärkein Lebesgue-Stieltjes–mitan erikoistapaus on *Lebesgue-mitta*, joka saadaan ottamalla edellä $F(x) = x$. $\mathbb{R}$:n Lebesgue-mitta mittaa siis välin pituuden. Yleisemmin Lebesgue-mitan $\mathbb{R}^n$:ssä voidaan intuitiivisesti ajatella mittaavan joukon $E \in \mathbb{R}^n$ "n-dimensiosta tilavuutta" (1-*d*: pituus, 2-*d*: pinta-ala, 3-*d*: "tavallinen" tilavuus, ...).

Kun μ on Borel-mitta $\mathbb{R}$:ssä s.e. $\mu(B) < \infty$ kaikilla rajoitetuilla Borel-joukoilla, voidaan määritellä oikealta jatkuva ei-vähenevä funktio G asettamalla $G(0) = 0$ ja

$$G(x) = \begin{cases} \mu((0, x]), & \text{kun} \quad x > 0, \\ -\mu((x, 0]), & \text{kun} \quad x < 0. \end{cases}$$

Tällöin $\mu = \mu_G$. Lebesgue-Stieltjes–mitta antaa siis kaikki Borel-mitat, jotka ovat äärellisiä rajoitetuilla joukoilla.

Määritelmä 1.13 (Tulomitta ja tulomitta-avaruus) *Olkoon $(X, \mathcal{M}, \mu)$ ja $(Y, \mathcal{N}, \nu)$ kaksi mitta-avaruutta. Näiden* tulomitta-avaruus *on*

$$(X \times Y, \mathcal{M} \otimes \mathcal{N}, \mu \times \nu),$$

missä $X \times Y$ on karteesinen tuloavaruus, $\mathcal{M} \otimes \mathcal{N}$ tensoritulo-σ-algebra, ja $\mu \times \nu$ tulomitta. Tulomitta $\mu \times \nu$ on sellainen mitta σ-algebralla $\mathcal{M} \otimes \mathcal{N}$, jolle pätee

$$\mu \times \nu(A \times B) = \mu(A)\nu(B)$$

kaikilla mitallisilla suorakulmioilla $A \times B$, $A \in \mathcal{M}$, $B \in \mathcal{N}$.

Mitta $\mu \times \nu$ on hyvin määritelty. Jos mitta-avaruudet ovat σ-äärellisiä, tulomitta on yksikäsitteinen, ja se on myös σ-äärellinen. Tulomittakonstruktio yleistyy luonnollisella tavalla n-kertaisiin tuloihin

$$\left(\prod_{i=1}^{n} X_i, \bigotimes_{i=1}^{n} \mathcal{M}_i, \prod_{i=1}^{n} \mu_i \right).$$

Esimerkiksi euklidisen avaruuden $\mathbb{R}^n$ Borel-mitta saadaan $\mathbb{R}$:n Borel-mitan (kopioiden) n-kertaisena tulona.

Lebesgue-integraali. Olkoon $(X, \mathcal{M}, \mu)$ kiinnitetty mitta-avaruus. Lebesgue-mitan μ avulla voidaan määritellä mitallisen funktion $f : X \to \mathbb{R}$ (tai $f : X \to$

$\bar{\mathbb{R}}$) *Lebesgue-integraali*, merk. $\int f d\mu$. (Sigma-algebrana on Borel-σ-algebra.) Lebesgue-integraalin konstruoiminen etenee vaiheissa siten, että integraali määritellään ensin ei-negatiivisille yksinkertaisille funktioille, sitten yleisille $f \geq 0$ monotonisen konvergenssin nojalla, ja lopulta yleisille $f = f^+ - f^-$ lineaarisuuden perusteella (f^+ ja f^- tarkoittavat f:n positiivisia ja negatiivisia osia).

Integraalille $\int f d\mu$ käytetään tilanteesta riippuen eri merkintätapoja, mm.

$$\int_A f(x)\mu(dx) \quad \text{tai} \quad \int_A f(x)d\mu(x)$$

kun integroidaan osajoukon $A \subset X$ yli ja integrointimuuttuja kirjoitetaan eksplisiittisesti esiin. Koulumatematiikassa opittu f:n Riemann-integraali määritellään Riemann-summien raja-arvona (mikäli se on olemassa), kun integrointivälin $[a, b]$ ositus kasvaa äärettömän tiheäksi. Koska olennaisesti f on Lebesgue-integroituva aina jos se on Riemann-integroituva, Riemann-integraalin merkintää käytetään joskus myös Lebesgue-integraalille reaaliakselilla, ja kirjoitetaan

$$\int_a^b f(x)dx \quad \text{eikä} \quad \int_{[a,b]} f d\mu,$$

vaikka f ei olisikaan Riemann-integroituva.

1.1.2 Todennäköisyysteoriaa

Esitetään seuraavaksi kertauksenomaisesti joitakin todennäköisyysteorian käsitteitä.

Todennäköisyysavaruus, satunnaismuuttujat. Todennäköisyyden perustukset ovat suoraan peräisin mittateoriasta. Monille käsitteille ja merkinnöille on kuitenkin omat vakiintuneet vastineensa todennäköisyysteoriassa. *Todennäköisyysavaruus* tai *todennäköisyyskenttä* $(\Omega, \mathscr{F}, \mathbb{P})$ on mitta-avaruus kokonaismassalla $\mathbb{P}(\Omega) = 1$. Kolmikossa $(\Omega, \mathscr{F}, \mathbb{P})$

- Ω on *otosavaruus* eli perusjoukko, jonka elementteinä ovat *alkeistapaukset* $\omega \in \Omega$.
- $\mathscr{F}$ on σ-algebra Ω:lla. $\mathscr{F}$:n mitallisia joukkoja kutsutaan *tapahtumiksi*.
- $\mathbb{P}$ on *todennäköisyysmitta*, eli mitta, jolle pätee $\mathbb{P}(\Omega) = 1$. $\mathbb{P}$ liittää todennäköisyyden jokaiseen $\mathscr{F}$:n joukkoon.

Satunnaismuuttuja on mitallinen kuvaus $X : (\Omega, \mathscr{F}) \to (E, \mathscr{E})$, joka ottaa arvonsa jossakin mitallisessa avaruudessa E. Yleensä $E = \mathbb{R}$; jos $E = \mathbb{R}^d$, X:ää kutsutaan satunnaisvektoriksi, ja jos E on funktioavaruus, X on satunnainen funktio. Osiossa 1.6 tavataan esimerkiksi pisteprosessien avaruus $M_p(E)$, jonka elementit ovat pisteprosesseja E:llä. σ-algebra $\mathscr{E}$ otetaan sopivaksi Borel-σ-algebraksi $\mathscr{B}$; esimerkiksi $\mathbb{R}$:n tapauksessa $(E, \mathscr{E}) = (\mathbb{R}, \mathscr{B}_{\mathbb{R}})$.

Kun $\omega \in \Omega$ on kiinnitetty, $X = X(\omega)$ tulkitaan satunnaismuuttujan (vektorin, ...) realisaatioksi, ja on siis reaaliluku (reaalilukuarvoinen vektori, ...).

Jakauma, kertymäfunktio ja tiheysfunktio. Satunnaismuuttujan X *jakauma* P on todennäköisyysmitta, joka saadaan kun todennäköisyysmitta $\mathbb{P}$ "siirretään" reaaliakselille,

$$P(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}\left(\omega \in \Omega \mid X(\omega) \in B\right)$$

kaikilla $B \in \mathscr{B}_{\mathbb{R}}$.

X:n *kertymäfunktio* $F : \mathbb{R} \to \mathbb{R}$ määritellään

$$F(x) = \mathbb{P}(X \leq x) = P((-\infty, x]).$$

Jakauma P on siis F:n Lebesgue-Stieltjes-mitta.

Funktio $f : \mathbb{R} \to \mathbb{R}$ on satunnaismuuttujan X *tiheysfunktio*, jos

$$F(x) = \int_{-\infty}^{x} f(s)\,\mathrm{d}s$$

kaikilla $x \in \mathbb{R}$.[1] Tällöin X:n jakaumaa sanotaan *jatkuvaksi*. Jos taas X keskittyy äärelliseen tai korkeintaan numeroituvaan joukkoon, ts. jos on olemassa reaaliluvat $x_1, x_2, \ldots$ s.e.

$$\mathbb{P}\left(X \in \{x_1, x_2, \ldots\}\right) = 1,$$

kutsutaan X:n jakaumaa *diskreetiksi*. Tässä tapauksessa jakauman *pistetodennäköisyysfunktio* (tai *todennäköisyysfunktio*) $p : \mathbb{R} \to \mathbb{R}$ määritellään ehdosta

$$p(x) = \mathbb{P}(X = x).$$

Odotusarvo. Reaaliarvoisen satunnaismuuttujan X *odotusarvo* on sen Lebesgue-integraali todennäköisyysmitan suhteen yli koko todennäköisyysavaruuden:

$$\mathbb{E}(X) = \int_{\Omega} X\,\mathrm{d}\mathbb{P}(\omega), \tag{1.1}$$

edellyttäen, että $\mathbb{E}(|X|) < \infty$. Jos $X \geq 0$ melkein varmasti (todennäköisyydellä 1), odotusarvoksi sallitaan myös ∞, jolloin odotusarvo on määritelty kaikille ei-negatiivisille satunnaismuuttujille.

Jos $h : \mathbb{R} \to \mathbb{R}$ on mielivaltainen rajoitettu mitallinen funktio, niin $h(X)$ (eli $h \circ X$) on myös satunnaismuuttuja, ja

$$\mathbb{E}(h(X)) = \int_{\Omega} h(X)\,\mathrm{d}\mathbb{P}(\omega) = \int_{\mathbb{R}} h(x)P(dx).$$

Kun X:n kertymäfunktio on F, voidaan odotusarvo kirjoittaa edelleen muodossa

$$\mathbb{E}(h(X)) = \int_{-\infty}^{\infty} h(x)\,\mathrm{d}F(x),$$

ja jos lisäksi X:llä on tiheysfunktio f, saadaan

$$\mathbb{E}(h(X)) = \int_{-\infty}^{\infty} h(x)f(x)\,\mathrm{d}x.$$

[1] Toisin sanoen, X:n tiheysfunktio on X:n jakauman derivaatta Lebesgue-mitan suhteen, $f = dP/dx$.

Viimeinen on muoto, jota käyttäen useimmat odotusarvot käytännössä lasketaan. Jos X:n jakauma on diskreetti ja todennäköisyysfunktiona on p, saadaan vastaavasti

$$\mathbb{E}(h(X)) = \sum_{i=1}^{\infty} h(x_i)p(x_i),$$

missä vaatimuksena on luonnollisesti, että $\sum_{i=1}^{\infty} p(x_i) = 1$.

Edellä esitetyt käsitteet yleistyvät luonnollisesti $\mathbb{R}^d$-arvoisiin satunnaisvektoreihin.

Satunnaismuuttujien yhtäsuuruudesta. Samalla otosavaruudella Ω määritellyt satunnaismuuttujat X ja Y ovat yhtäsuuria funktioina, jos $X(\omega) = Y(\omega)$ kaikilla $\omega \in \Omega$. Tämä on kuitenkin käytännössä yleensä liian vahva vaatimus, sillä se vaatii yhtäsuuruutta myös kaikilla nollamittaisilla joukoilla (joukoilla, joiden todennäköisyys on nolla). Hyödyllisempi käsite on *melkein varma* (m.v.) yhtäsuuruus: $X = Y$ m.v., jos $\mathbb{P}(X = Y) = 1$.[2] X ja Y ovat *samoin jakautuneita*, jos $\mathbb{P}(X \in B) = \mathbb{P}(Y \in B)$ kaikille mitallisille joukoille B X:n ja Y:n (yhteisessä) arvojoukossa eli kuvassa. Satunnaismuuttujien samoinjakautuneisuudesta voidaan puhua, vaikka ne olisi määritelty eri todennäköisyysavaruuksissa. Samoinjakautuneisuutta merkitään $X \overset{d}{=} Y$ tai $X =_L Y$ (in distribution, in law).

Informaatiosta ja riippumattomuudesta. Todennäköisyysteoriassa σ-algebrat edustavat informaatiota, ja niiden mitalliset osajoukot ovat tapahtumia: Todennäköisyysavaruuden $(\Omega, \mathscr{F}, \mathbb{P})$ σ-algebra $\mathscr{F}$ edustaa kaikkea informaatiota, kun taas ali-σ-algebra $\mathcal{G} \subset \mathscr{F}$ edustaa osittaista informaatiota. Kun "tiedetään" σ-algebra $\mathcal{G}$, tiedetään kaikille tapahtumille $A \in \mathcal{G}$ tapahtuiko A vai ei.

Jos X on satunnaismuuttuja Ω:lla, on X:n *generoima* σ-algebra

$$\sigma(X) = \left\{ X^{-1}(B) \,\middle|\, B \in \mathscr{B}_{\mathbb{R}} \right\} = \{\{X \in B\} \mid B \in \mathscr{B}_{\mathbb{R}}\}$$

(vrt. edellinen alaosio). X:n mitallisuus on yhtäpitävää ehdon $\sigma(X) \subseteq \mathscr{F}$ kanssa. Jos X:n arvo tiedetään, on tämä sama kuin tiedettäisiin – kaikilla $\{B \in \mathscr{B}_{\mathbb{R}}\}$ – tapahtuiko $\{X \in B\}$; X voi kuitenkin luonnollisesti saada saman arvon $X = X(\omega)$ useilla $\omega \in \Omega$, joten tieto X:stä ei riitä määrittämään, mikä tulemista ω itse asiassa tapahtui. Tässä mielessä $\sigma(X)$ edustaa osittaista informaatiota.

Kaksi tapahtumaa A ja B ovat riippumattomia, jos $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. A:n ehdollinen todennäköisyys ehdolla B määritellään

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

kun $\mathbb{P}(B) > 0$. Riippumattomuus voidaan siis vaihtoehtoisesti ilmaista ehdon $\mathbb{P}(A|B) = P(A)$ kautta: intuitiivisesti, tieto B:stä ei muuta A:n todennäköisyyttä. Tämä yksinkertainen riippumattomuuden määritelmä ei kuitenkaan riitä kaikissa tilanteissa. Yleinen määritelmä riippumattomuudelle annetaan σ-algebrojen kautta:

[2] Analyysissa käytetään terminologiaa (μ-)*melkein kaikkialla*, todennäköisyysteoriassa puolestaan ($\mathbb{P}$-)*melkein varmasti*.

Määritelmä 1.14 (σ-algebrojen riippumattomuus)
Olkoon $\{\mathcal{G}_i\}_{i=1}^n$ kokoelma $\mathscr{F}$:n ali-σ-algebroja. $\mathcal{G}_1, \ldots, \mathcal{G}_n$ ovat (keskenään) riippumattomia*, jos millä tahansa tapahtumilla $A_1, \ldots, A_n$, $A_1 \in \mathcal{G}_1, \ldots, A_n \in \mathcal{G}_n$,*

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n). \tag{1.2}$$

Mielivaltainen $\mathscr{F}$:n ali-σ-algebrojen kokoelma $\{\mathcal{G}_i : i \in I\}$ on riippumaton, jos sen jokainen äärellinen alikokoelma on riippumaton.

Satunnaismuuttujien ja tapahtumien riippumattomuus määritellään edelliseen perustuen.

Määritelmä 1.15 (Satunnaismuuttujien riippumattomuus)
Satunnaismuuttujien $\{X_i : i \in I\}$ kokoelma todennäköisyysavaruudella $(\Omega, \mathscr{F}, \mathbb{P})$ on riippumaton*, jos (ja vain jos) satunnaismuuttujien generoimien σ-algebrojen kokoelma*

$$\{\sigma(X_i) : i \in I\}$$

on. Yhtäpitävästi, olkoon $(i_j)_{j=1}^n$ mikä tahansa äärellinen joukko erillisiä ($i_j \neq i_k, j \neq k$) indeksejä, ja $\{B_i\}_{i=1}^n$ mitä tahansa mitallisia joukkoja satunnaismuuttujien maalijoukoissa: jos

$$\mathbb{P}\left(X_{i_1} \in B_1, \ldots, X_{i_n} \in B_n\right) = \prod_{j=1}^{n} \mathbb{P}\left(X_{i_j} \in B_j\right), \tag{1.3}$$

niin satunnaismuuttujat $\{X_i : i \in I\}$ ovat riippumattomia, ja päinvastoin.

Edelleen tapahtumat $\{A_i : i \in I\}$ ovat riippumattomia, jos vastaavat indikaattorisatunnaismuuttujat $\{\mathbb{1}_{A_i} : i \in I\}$ ovat.

Riippumattomuudella on liittymäkohtia tulomitan käsitteeseen. Olkoon esimerkiksi P satunnaisvektorin $X = (X_1, \ldots X_n) \in \mathbb{R}^n$ jakauma, ja olkoon P_i komponentin $X_i \in \mathbb{R}$ jakauma, $i = 1, \ldots, n$. Satunnaismuuttujat $X_1, \ldots, X_n$ ovat riippumattomia jos ja vain jos mitalle P pätee $P = P_1 \otimes \cdots \otimes P_n$.

Voidaan osoittaa, että riippumattomuuden näyttämiseksi riittää osoittaa ominaisuudet (1.2) tai (1.3) sellaisten joukkojen luokille, jotka ovat suljettuja leikkausten suhteen ja generoivat kyseessä olevat σ-algebrat. Täten saadaan tuttu ehto riippumattomuudelle kertymäfunktioiden kautta, esimerkiksi

$$\mathbb{P}\left(X_{i_1} \leq x_1, \ldots, X_{i_n} \leq x_n\right) = \prod_{j=1}^{n} \mathbb{P}\left(X_{i_j} \leq x_j\right),$$

kun kukin komponentti X_i ottaa arvonsa $\mathbb{R}$:ssä. Edelleen, jos esimerkiksi satunnaisvektorilla (X, Y) on olemassa tiheysfunktio $\mathbb{R}^2$:lla, niin X ja Y ovat riippumattomia jos ja vain jos $f(x, y) = f_X(x) f_Y(y)$, eli tiheysfunktio faktorisoituu (f_X ja f_Y ovat X:n ja Y:n reunajakaumienn tiheysfunktioita).

Ehdollinen odotusarvo ja ehdollistaminen. Satunnaismuuttuja X kuuluu avaruuteen $L^1(\mathbb{P}) = L^1(\Omega, \mathscr{F}, \mathbb{P})$, jos $\mathbb{E}(X) = \int_\Omega |X(\omega)| \, \mathrm{d}\mathbb{P}(\omega) < \infty$.

Määritelmä 1.16 (Ehdollinen odotusarvo)
Olkoon $(\Omega, \mathscr{F}, \mathbb{P})$ todennäköisyysavaruus, $X \in L^1(\mathbb{P})$ ja σ-algebra $\mathcal{G}$ $\mathscr{F}$:n ali-σ-algebra. Satunnaismuuttujan X ehdollinen odotusarvo *$\mathcal{G}$:n suhteen, merk.*

$\mathbb{E}(X|\mathcal{G})$, *on satunnaismuuttuja* $Y \in \mathcal{G}$ *jolle pätee*

$$\int_G X \, \mathrm{d}\mathbb{P} = \int_G Y \, \mathrm{d}\mathbb{P}, \quad \forall\, G \in \mathcal{G}. \tag{1.4}$$

Ehdollinen odotusarvo on yksikäsitteinen melkein varmasti, ts. jos Z on toinen $\mathcal{G}$-mitallinen satunnaismuuttuja jolle (1.4) pätee, niin $\mathbb{P}(Y = Z) = 1$. Jos satunnaismuuttujat ymmärretään satunnaismuuttujien ekvivalenssiluokkina siten että melkein varmasti (todennäköisyydellä 1) yhtäsuuret satunnaismuuttujat samastetaan keskenään, kuten yleensä on tapana, on ehdollinen odotusarvo siis yksikäsitteinen.

Listataan joitakin ehdollisen odotusarvon ominaisuuksia.

Propositio 1.17 (Ehdollisen odotusarvon ominaisuuksia)
Olkoon $X, Z \in L^1(\Omega, \mathscr{F}, \mathbb{P})$ *ja* $\mathcal{G} \subset \mathscr{F}$. *Asetetaan* $Y = \mathbb{E}(X|\mathcal{G})$.

1. $\mathbb{E}(X) = \int_\Omega X \, \mathrm{d}\mathbb{P} = \int_\Omega Y \, \mathrm{d}\mathbb{P} = \mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(X|\mathcal{G}))$, *siis* $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$.
2. *Jos* X *on* $\mathcal{G}$*-mitallinen eli* $X \in \mathcal{G}$, *niin* $\mathbb{E}(X|\mathcal{G}) = X$.
3. *Jos* X *on riippumaton* $\mathcal{G}$*:stä (eli* $\sigma(X)$ *ja* $\mathcal{G}$ *ovat riippumattomia), niin* $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$.
4. *Jos* $Z \in \mathcal{G}$ *ja* $XZ \in L^1(\mathbb{P})$, *niin* $\mathbb{E}(ZX|\mathcal{G}) = Z\mathbb{E}(X|\mathcal{G})$.
5. *Jos* $\mathcal{H} \subset \mathcal{G}$, *niin* $\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(X|\mathcal{H})$.
6. *Jos* $\mathcal{H} \subset \mathcal{G}$ *ja* $\mathbb{E}(X|\mathcal{G}) \subset \mathcal{H}$, *niin* $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X|\mathcal{H})$.
7. $\mathbb{E}(aX + bZ + c|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(X|\mathcal{G}) + c$, $a, b, c \in \mathbb{R}$, *eli ehdollinen odotusarvo on lineaarinen.*
8. *Jos* $X \geq Z$ *niin* $\mathbb{E}(X|\mathcal{G}) \geq \mathbb{E}(Z|\mathcal{G})$.
9. Jensenin epäyhtälö: *Jos* $g : \mathbb{R} \to \mathbb{R}$ *on konveksi ja* $g(X) \in L^1(\mathbb{P})$, *niin* $g(\mathbb{E}(X|\mathcal{G}) \leq \mathbb{E}(g(X)|\mathcal{G})$.

Ehdollinen todennäköisyys määritellään ehdollisen odotusarvon avulla: olkoon $X = \mathbb{1}_B$ tapahtuman B indikaattorisatunnaismuuttuja, jolloin voidaan kirjoittaa $\mathbb{P}(B|\mathcal{G}) = \mathbb{E}(\mathbb{1}_B|\mathcal{G})$. Kun σ-algebra, jonka suhteen ehdollistetaan, on satunnaismuuttujan Y generoima, $\mathcal{G} = \sigma(Y)$, kirjoitetaan tavallisesti $E(X|Y)$ pidemmän $E(X|\sigma(Y))$ sijasta.

Satunnaismuuttujien suppenemisesta. Esitetään luettelonomaisesti satunnaismuuttujia koskevia suppenemiskäsitteitä.

Määritelmä 1.18 (Satunnaismuuttujien suppeneminen)
Olkoon $(X_n : n \geq 1)$, X *reaaliarvoisia satunnaismuuttujia.*

(i) Melkein varma suppeneminen. $X_n \to X$ *melkein varmasti, jos*

$$\mathbb{P}\left(\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right) = 1.$$

(ii) Todennäköisyyden suhteen suppeneminen. $X_n \to X$ *todennäköisyyden suhteen jos, kaikilla* $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon\right) = 0.$$

(iii) L^p-suppeneminen. $X_n \to X$ L^p*:ssä,* $1 \leq p < \infty$, *jos*

$$\lim_{n\to\infty} \mathbb{E}\left(|X_n(\omega) - X(\omega)|^p\right) = 0.$$

(iv) Jakauman suhteen suppeneminen *eli* heikko suppeneminen. $X_n \to X$ *jakauman suhteen (heikosti), jos*

$$\lim_{n\to\infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x)$$

kaikilla $x \in \mathbb{R}$ *joissa* $F(x) = \mathbb{P}(X \leq x)$ *on jatkuva.*

Huomautetaan, että kolme ensimmäistä suppenemismoodia edellyttävät että satunnaismuuttujat on määritelty samalla todennäköisyysavaruudella. Jakauman suhteen suppeneminen sen sijaan ei edellytä tätä.

Yllä esitetty heikon suppenemisen määritelmä satunnaismuuttujille $\mathbb{R}$:ssä on yleisen määritelmän erikoistapaus. Olkoon $\{\mu_n\}$, μ Borel-todennäköisyysmittoja metrisellä avaruudella E. Mitta $\mu_n \to \mu$ heikosti, jos

$$\int_E f d\mu_n \to \int_E f d\mu,$$

kaikilla E:n jatkuvilla, rajoitetuilla funktioilla f. Satunnaismuuttujat suppenevat heikosti, jos niiden jakaumat suppenevat esitetyssä mielessä.

Stokastiset prosessit. Olkoon $(\Omega, \mathscr{F}, \mathbb{P})$ todennäköisyysavaruus ja $(E, \mathscr{E})$ mitallinen avaruus. E-arvoinen *stokastinen prosessi* yleisesti on kokoelma $\{X_\alpha : \alpha \in I\}$ E:ssä arvonsa ottavia satunnaismuuttujia määriteltynä samassa otosavaruudessa Ω, missä I on mielivaltainen indeksijoukko. Avaruutta E kutsutaan prosessin *tila-avaruudeksi.*

Tyypillisesti stokastinen prosessi kuvaa jonkin satunnaisilmiön kehitystä ajan suhteen, jolloin prosessin indeksijoukko edustaa aikaa, ollen yleensä $\mathbb{R}_+ = [0, \infty)$ tai sen osajoukko. Tila-avaruus on tyypillisesti euklidinen avaruus $\mathbb{R}^d$, $d \geq 1$, tai jokin sen osajoukko. Luonnollinen σ-algebra E:llä on sen Borel-σ-algebra, $\mathscr{E} = \mathscr{B}_E$.

Olkoon $X = \{X_t : t \in [0, \infty)\}$ stokastinen prosessi, jolloin $X_t = X_t(\omega)$ on siis satunnaismuuttuja jokaisella $0 \leq t < \infty$. X voidaan tulkita funktioksi $\Omega \times \mathbb{R}_+ \to E$ ottamalla $X(\omega, t) = X_t(\omega)$. Jokaisella $\omega \in \Omega$ funktiota $t \mapsto X_t(\omega)$ kutsutaan prosessin X *poluksi* tai realisaatioksi.

Satunnaismuuttujien yhteydessä edellä mainittiin, että σ-algebrat edustavat informaatiota todennäköisyysavaruuden tapahtumista. Stokastisten prosessien yhteydessä on luonnollista tarkastella informaation kehitystä (lisääntymistä) ajassa. Seuraava käsite määrittelee tämän formaalisti.

Määritelmä 1.19 (Filtraatio)
Filtraatio *todennäköisyysavaruudella* $(\Omega, \mathscr{F}, \mathbb{P})$ *on kokoelma* σ*-algebroja* $\{\mathscr{F}_t : t \in \mathbb{R}_+\}$, *jolle pätee*

$$\mathscr{F}_s \subseteq \mathscr{F}_t \subseteq \mathscr{F}, \quad \forall\, 0 \leq s < t < \infty.$$

Esimerkki 1.20 (Prosessin historia)
Olkoon $X = (X_t)_{t\geq 0}$ *stokastinen prosessi.* X*:n generoimaa* σ*-algebraa* $\mathscr{F}_t^X = \sigma(X_s : 0 \leq s \leq t)$ *kutsutaan usein prosessin* historiaksi. $\mathscr{F}_t^X$ *sisältää siis hetkeen* t *mennessä "havaitun" informaation prosessin* X *kehityksestä.*

Prosessin $X = (X_t)$ sanotaan olevan *sovitettu* (adapted) filtraatioon $(\mathscr{F}_t)$, jos X_t on $\mathscr{F}_t$-mitallinen jokaisella $t \in [0, \infty)$. Pienin filtraatio johon X on sovitettu on sen itsensä generoima filtraatio $\mathscr{F}_t^X$. Prosessi X on *mitallinen*, jos se on $\mathscr{F} \otimes \mathscr{B}_{\mathbb{R}}$-mitallinen funktiona $X : \Omega \times \mathbb{R} \to E$.

1.2 Satunnaismuuttujajonojen maksimeista

Olkoon $X_1, X_2, \ldots$ jono riippumattomia ja samoin jakautuneita (independent and identically distributed, iid) ei-degeneroituneita[3] satunnaismuuttujia kiinnitetyllä todennäköisyyskentällä $(\Omega, \mathscr{F}, \mathbb{P})$, ja F näiden yhteinen kertymäfunktio. (Riippumattomuusoletuksesta tullaan myöhemmin luopumaan.) Satunnaismuuttujat $X_i : (\Omega, \mathscr{F}) \to (E, \mathscr{E})$ ottavat arvonsa mitallisessa avaruudessa $(E, \mathscr{E})$: tässä esityksessä on $E = \mathbb{R}^d$ tai sen osajoukko, $d \geq 1$, varustettuna vastaavalla Borel-sigma-algebralla $\mathscr{B}$. Jatkossa ilman eri mainintaa $E = \mathbb{R}$.

Ääriarvoteoriassa kiinnostuksen kohteena on *otosmaksimien*

$$M_1 = X_1, \quad M_n = \max(X_1, \ldots, X_n), \quad n \geq 2,$$

käyttäytyminen.[4] Sovelluksissa satunnaismuuttujat X_i edustavat yleensä tasavälisiä havaintoja tarkasteltavasta prosessista, kuten esimerkiksi merenpinnan korkeuden tuntihavaintoja tai osakkeen päivittäisiä tuottoja.

Periaatteessa otosmaksimin M_n täsmällinen jakauma voidaan johtaa tarkasti kaikilla n:

$$\begin{aligned} \mathbb{P}(M_n \leq x) &= \mathbb{P}(X_1 \leq x, \ldots, X_n \leq x) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i \leq x) = F^n(x), \ x \in \mathbb{R}, \ n \in \mathbb{N}. \end{aligned} \tag{1.5}$$

Tämä ei kuitenkaan ole erityisen hyödyllinen esitys käytännön kannalta, sillä jakauman F tarkkaa muotoa ei käytännön sovelluksissa useinkaan tunneta. Eräs mahdollisuus on edetä estimoimalla F havaitusta aineistosta perinteisiä tilastomenetelmiä hyödyntäen, ja käyttää saatua estimaattia $\hat{F}$ yllä.

[3] Tällä tarkoitetaan satunnaismuuttujia, joiden jakauma ei ole keskittynyt yhteen pisteeseen. Samaa termiä käytetään myös jakaumista tai kertymäfunktioista: kertymäfunktio F on ei-degeneroitunut, jos se ei ole muotoa $F(x) = \mathbb{1}_{\{x>c\}}$ jollain $c \in \mathbb{R}$.

[4] Tässä esityksessä keskitytään tarkastelemaan maksimeja, mutta vastaavat tulokset minimeille seuraavat suoraan käyttämällä relaatiota $\min(X_1, \ldots, X_n) = -\max(-X_1, \ldots, -X_n)$.

Toinen lähestymistapa on hyväksyä että F on tuntematon, ja sen sijaan etsiä suoraan mallia maksimien jakaumalle F^n. Tämä osoittautuu erittäin tulokselliseksi myös käytännön sovellusten kannalta.

On siis tarpeen tutkia tarkemmin, mitä otosmaksimin jakaumasta voidaan sanoa yleisessä tapauksessa. Satunnaisilmiöiden ääriarvot osuvat jakauman häntiin, otosmaksimien osalta oikeaan häntään: intuitiivisesti on selvää, että suureen M_n asymptoottisen käyttäytymisen (kun $n \to \infty$) täytyy olla yhteydessä jakauman F häntään lähellä jakauman oikeaa päätepistettä

$$x_F = \sup \{x \in \mathbb{R} \mid F(x) \leq 1\}.$$

Tarkastellessa jakauman F^n käytöstä kun $n \to \infty$ saadaan kuitenkin välittömästi

$$\begin{aligned} &\mathbb{P}(M_n \leq x) = F^n(x) \overset{n\to\infty}{\longrightarrow} 0, \quad && x \leq x_F \leq \infty, \\ &\mathbb{P}(M_n \leq x) = F^n(x) = 1, && x \geq x_F, x_F \leq \infty. \end{aligned}$$

Siis otosmaksimin M_n todennäköisyysmassa keskittyy pisteeseen x_F, kun $n \to \infty$.[5] Tämä ei tarjoa vielä lisäinformaatiota jakaumasta.

Osoittautuu, että enemmän tietoa maksimien käyttäytymisestä saadaan tarkastelemalla keskitettyjen ja normalisoitujen maksimien heikkoa suppenemista (eli suppenemista jakauman suhteen, in distribution). Tämä on klassisen ääriarvoteorian pääaiheita. Merkitään otosmaksimin M_n affiinia muunnosta

$$M_n^* = \frac{M_n - d_n}{c_n},$$

missä (c_n) ja (d_n) ovat jono vakioita. Sopivat normeerausvakioiden valinnat "stabiloivat" M_n^*:n jakauman lokaation ja skaalan n:n kasvaessa, jolloin vältetään normeeraamattoman otosmaksimin tarkastelussa kohdatut ongelmat. Seuraavassa luvussa esitetään Fisherin ja Tippettin mukaan nimetty keskeinen lause, joka olennaisesti kertoo, että mikäli on olemassa sellainen normeeraus että M_n^* yllä suppenee heikosti johonkin (ei-degeneroituneeseen) jakaumaan kun $n \to \infty$, tämän jakauman täytyy olla tyypiltään yksi kolmesta ns. standardista *ääriarvojakaumasta*.

Tulos voidaan nähdä ääriarvoteorian analogiaksi keskeiselle raja-arvolauseelle: Olettaen että satunnaismuuttujat $X_1, X_2, \ldots$ ovat riipumattomia ja samoin jakautuneita ja niiden varianssi on äärellinen, keskeinen raja-arvolause standardimuodossaan sanoo, että sopivasti normalisoitujen summien $S_n = X_1 + \ldots + X_n$ ainoa mahdollinen rajajakauma on normaalijakauma:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - a_n}{b_n}\right) = \Phi(x), \quad x \in \mathbb{R},$$

missä $a_n = n\mathbb{E}(X_i)$ ja $b_n = \sqrt{\mathrm{Var}(X_i)}$. Yleisemmin, iid satunnaismuuttujien normalisoitujen summien mahdolliset rajajakaumat kuuluvat stabiilien jakaumien (stable distributions) perheeseen, joiden erikoistapaus on normaalijakauma.

[5] Täsmällisesti, esityksestä seuraa, että $M_n \xrightarrow{P} x_F$ kun $n \to \infty$, missä siis $x_F \leq \infty$ (P edellä tarkoittaa todennäköisyyden suhteen suppenemista). Koska jono (M_n) on ei-vähenevä n:n suhteen, se suppenee melkein varmasti (m.v.), ja pätee myös $M_n \xrightarrow{\text{m.v.}} x_F$.

Siinä missä keskeinen raja-arvolause koskee iid satunnaismuuttujien summia, ääriarvoteorian "vastine" koskee satunnaismuuttujien *maksimeja*.

1.3 Ääriarvojakaumat ja yleistetty ääriarvojakauma

Seuraava lause on klassisen ääriarvoteorian perusta.

Lause 1.21 (Fisher-Tippett, Gnedenko; maksimien rajajakaumat) *Olkoon $(X_i)_{i=1}^n$ jono iid satunnaismuuttujia. Jos on olemassa normeerausvakioiden jonot (c_n) ja (d_n), $c_n \geq 0, d_n \in \mathbb{R}$, ja ei-degeneroitunut jakauma H siten, että*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - d_n}{c_n} \leq x\right) = \lim_{n\to\infty} F^n(c_n x + d_n) = H(x), \quad x \in \mathbb{R}, \tag{1.6}$$

niin H kuuluu yhteen seuraavista kolmesta jakaumaperheestä:

$$\textit{Fréchet:} \quad \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ \exp\{-x^{-\alpha}\}, & x \geq 0 \end{cases}, \qquad \alpha > 0,$$

$$\textit{Weibull:} \quad \Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^{-\alpha}\}, & x \leq 0 \\ 1, & x \geq 0 \end{cases}, \quad \alpha > 0,$$

$$\textit{Gumbel:} \quad \Lambda(x) = \exp\{-e^{-x}\}, \qquad x \in \mathbb{R}.$$

Yllä esiintyvät jakaumat tunnetaan kollektiivisesti ääriarvojakaumina. Jakaumat ovat lauseessa standardimuodossaan: kutakin tyyppiä vastaavat lokaatioskaala –perheet saadaan korvaamalla argumentti x yllä tekijällä $(x-\mu)/\sigma$, jolloin määrittelyalueiksi kahdessa ensimmäisessä tapauksessa tulee vastaavasti $x \leq \mu$ tai $x \geq \mu$. Fréchet- ja Weibull-jakaumien yhteydessä esiintyvä α on jakauman muodon määräävä parametri.

Lause 1.21 karakterisoi kaikki mahdolliset rajajakaumat: jos löytyy sopivat normeerausvakiot siten, että normeeratut otosmaksimit $c_n^{-1}(M_n - d_n)$ suppenevat jakauman suhteen, niin suppeneminen tapahtuu satunnaismuuttujaan, jonka jakauma on välttämättä yksi kolmesta lauseen ääriarvojakaumasta. Merkittävää tuloksessa on, että ääriarvojakaumat ovat ainoita mahdollisia rajajakaumia iid satunnaismuuttujien $X_1, X_2, \ldots$ otosmaksimeille M_n^* riippumatta itse muuttujien X_i jakaumasta F.

Huomautus 1.22

1) *Lauseen 1.21 rajajakauma ei aina ole olemassa, ks. esim. [2, s. 117]. Tällaisia tapauksia voidaan kuitenkin pitää poikkeuksina: käytännössä kaikilla normaalisti tavattavilla jakaumilla rajajakauma H on olemassa.*

2) *Rajajakauma* (1.6)*:ssä on yksikäsitteinen vain affiineja muunnoksia vaille. Jos jakauma esiintyy muodossa $H(cx + d)$,*

$$\lim_{n\to\infty} \mathbb{P}((M_n - d_n)/c_n \leq x) = H(cx + d),$$

niin $H(x)$ on myös raja-arvo normeerausvakioita muuttamalla:

$$\lim_{n\to\infty} \mathbb{P}((M_n - \hat{d}_n)/\hat{c}_n \leq x) = H(x),$$

missä $\hat{c}_n = c_n/c$ ja $\hat{d}_n = d_n/d$. Erityisesti siis normeerausvakiot on aina mahdollista valita niin, että rajajakauma H esiintyy standardimuodossaan.

Jatkoa ajatellen on hyödyllistä yhdistää lauseen 1.21 ääriarvojakaumat yhden jakaumaperheen alle ottamalla käyttöön parametrisaatio

$$H_\xi = \begin{cases} \Phi_{1/\xi}, & \xi > 0, \\ \Lambda, & \xi = 0, \\ \Psi_{-1/\xi}, & \xi < 0. \end{cases}$$

Jakauma H voidaan siis ajatella kolmen alla olevan ääriarvojakauman yleistyksenä, ja se tunnetaankin yleistettynä ääriarvojakaumana (generalized extreme value distribution, GEV). Määritelmä seuraa:

Määritelmä 1.23 (Yleistetty ääriarvojakauma, GEV)
Määritellään H_ξ ehdosta

$$H_\xi(x) = \begin{cases} \exp\left\{-(1+\xi x)^{-\frac{1}{\xi}}\right\}, & \xi \neq 0, \\ \exp\left\{-e^{-x}\right\}, & \xi = 0, \end{cases}$$

missä $1 + \xi x > 0$. Jakauma H on standardi yleistetty ääriarvojakauma parametrilla ξ. Vastaava lokaatio-skaala –perhe saadaan asettamalla $H_{\xi,\mu,\sigma}(x) := H_\xi((x-\mu)/\sigma)$ lokaatioparametrille $\mu \in \mathbb{R}$ ja skaalaparametrille $\sigma > 0$. $H_{\xi,\mu,\sigma}$:n määrittely-alue on

$$\begin{aligned} x > \mu - \tfrac{\sigma}{\xi}, &\quad \textit{kun} \quad \xi > 0, \\ x < \mu - \tfrac{\sigma}{\xi}, &\quad \textit{kun} \quad \xi < 0, \\ x \in \mathbb{R}, &\quad \textit{kun} \quad \xi = 0. \end{aligned}$$

Jakaumaa $H_{\xi,\mu,\sigma}$ kutsutaan yleistetyksi ääriarvojakaumaksi.

Parametri ξ on GEV-jakauman muotoparametri, ja H_ξ määrittelee jakauman tyypin, eli jakaumaperheen joka on kiinnitetty lokaatiota ja skaalausta lukuunottamatta.[6] Jakaumafunktio H_0 tapauksessa $\xi = 0$ tulkitaan H_ξ:n raja-arvona kun $\xi \to 0$: kiinteällä x pätee $\lim_{\xi\to 0} H_\xi(x) = H_0(x)$ (molemmilta puolilta), joten määritelmän 1.23 parametrisaatio on jatkuva ξ:n suhteen. GEV-parametrisaation jatkuvuus on erityisen hyödyllinen tilastollisia sovelluksia ajatellen; ks. osio 2.2.

Määritelmä 1.24 (Jakauman vaikutuspiiri maksimin suhteen, MDA)
Jos (1.6) *pätee jollekin ei-degeneroituneelle jakaumalle H, niin sanotaan, että F (vast. X) kuuluu jakauman H vaikutuspiiriin maksimin suhteen (F belongs to the maximum domain of attraction of H). Merkitään $F \in \mathrm{MDA}(H)$.*

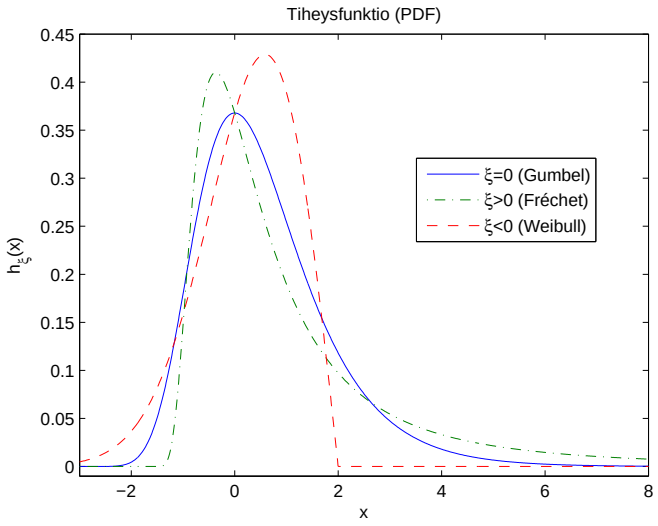
[6] Täsmällisemmin, kahden satunnaismuuttujan X ja Y (tai niiden jakaumien) sanotaan olevan samaa tyyppiä, jos on olemassa vakiot $a > 0$ ja $b \in \mathbb{R}$ s.e. $X \overset{d}{=} aY + b$.

Kuten ääriarvoteorian terminologian kohdalla usein, käsitteellä "maximum domain of attraction" ei ole vakiintunutta suomenkielistä vastinetta. Tässä esityksessä käytetään ilmaisua "vaikutuspiiri maksimin suhteen", joka on peräisin H. Nyrhiseltä [13].

Lause 1.21 voidaan yhtäpitävästi esittää määritelmää 1.24 käyttäen.

Lause 1.25 (Fisher-Tippett, Gnedenko)
Jos $F \in \mathrm{MDA}(H)$ jollekin ei-degeneroituneelle jakaumalle H, niin H:n täytyy olla tyyppiä H_ξ, eli yleistetyn ääriarvojakaumaperheen jäsen.

GEV-jakauma tapauksessa $\xi > 0$ on siis Fréchet-jakauma, tapauksessa $\xi = 0$ Gumbel-jakauma, ja tapauksessa $\xi < 0$ Weibull-jakauma. Jakauman tiheysfunktio eri tapauksia vastaten on esitetty kuvassa 1.1, ja kertymäfunktio vastaavasti kuvassa 1.2.

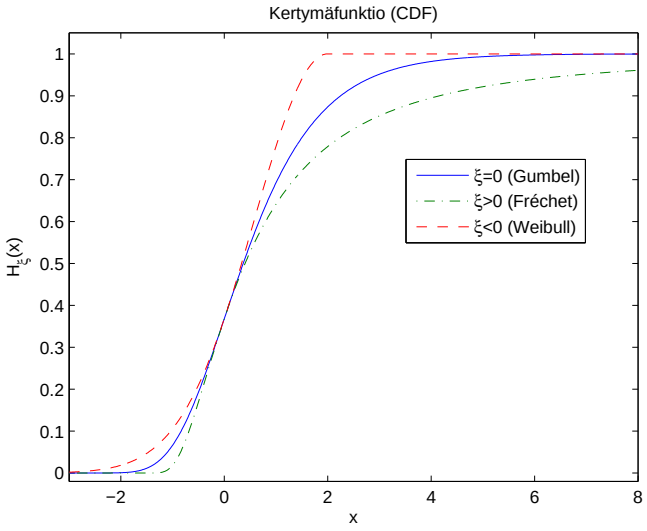


Kuva 1.1: GEV-jakauman tiheysfunktio.

Ääriarvojakaumista Weibull on lyhythäntäinen jakauma, jolla on äärellinen oikea päätepiste x_F. Fréchet- ja Gumbel –jakaumilla on ääretön oikea päätepiste, mutta Fréchet-jakauman häntätodennäköisyyksien vaimeneminen on paljon hitaampaa kuin Gumbel-jakauman: Fréchet'n vaikutuspiiriin maksimin suhteen kuuluvat jakaumat ovat paksuhäntäisiä, Gumbelin vaikutuspiiriin kuuluvat jakaumat puolestaan ohuthäntäisiä tai keskipaksuja.

Esitetään seuraavaksi esimerkkejä normeerauskertoimista ja vaikutuspiiristä tunnettujen jakaumien osalta.

Esimerkki 1.26 (Eksponenttijakauma)
Olkoon $X_1, X_2, \ldots$ iid jono eksponenttijakautuneita satunnaismuuttujia parametrilla $\mu > 0$ yhteisenä kertymäfunktiona F, $F(x) = 1 - e^{-\mu x}$, $x \geq 0$. Va-



Kuva 1.2: GEV-jakauman kertymäfunktio.

litsemalla normeerausvakioiden jonot $c_n = 1/\mu$ *ja* $d_n = \ln(n/\mu)$ *saadaan*

$$\begin{aligned}\mathbb{P}\left(\frac{M_n - d_n}{c_n} \leq x\right) &= F^n\left(c_n x + d_n\right) = F^n\left(\frac{x}{\mu} + \frac{\ln n}{\mu}\right) \\ &= \left(1 - e^{-(x - \ln n)}\right)^n = \left(1 - \frac{1}{n}e^{-x}\right)^n, \quad x \geq \ln n, \\ \lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - d_n}{c_n} \leq x\right) &= \exp\left(-e^{-x}\right), \quad x \in \mathbb{R}.\end{aligned}$$

Siis rajajakauma on Gumbel, $F \in \mathrm{MDA}(H_0)$*, vastaten tapausta* $\xi = 0$*.*

Esimerkki 1.27 (Tasajakauma)
Olkoon F *tasajakauma yksikkövälillä,* $F(x) = x$*,* $x \in (0, 1)$*. Valitaan* $c_n = 1/n$ *ja* $d_n = 1$*. Kiinteällä* $y < 0$

$$\begin{aligned}\mathbb{P}\left(\frac{M_n - d_n}{c_n} \leq y\right) &= F^n\left(\frac{y}{n} + 1\right) = \left(1 + \frac{y}{n}\right)^n, \quad y > -n, \\ \lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - d_n}{c_n} \leq y\right) &= e^{-y}, \quad y \in \mathbb{R}.\end{aligned}$$

Rajajakauma on siis yleistetty ääriarvojakauma parametrilla $\xi = -1$ *eli Weibull,* $F \in \mathrm{MDA}(H_{-1})$*.*

Esimerkki 1.28 (Pareto-jakauma)
Olkoon F *Pareto-jakauma* $\mathrm{Pa}(\alpha, \kappa)$*,* $F(x) = 1 - (\kappa/(\kappa + x))^\alpha$*,* $\alpha > 0$*,* $\kappa > 0$*,*

$x \geq 0$. *Asettamalla* $c_n = \kappa n^{1/\alpha}/\alpha$ *ja* $d_n = \kappa n^{1/\alpha} - \kappa$, *saadaan*

$$\begin{aligned}
\mathbb{P}\left(\frac{M_n - d_n}{c_n} \leq x\right) &= F^n\left(\frac{\alpha\left(x - \kappa n^{1/\alpha} + \kappa\right)}{\kappa n^{1/\alpha}}\right) \\
&= \left(1 - \left(\frac{\kappa}{\kappa + \kappa n^{1/\alpha} x/\alpha + \kappa n^{1/\alpha} - \kappa}\right)^{\alpha}\right)^n \\
&= \left(1 - \left(\frac{1}{n^{1/\alpha}\left(1 + x/\alpha\right)}\right)^{\alpha}\right)^n \\
&= \left(1 - \frac{1}{n}\left(1 + \frac{x}{\alpha}\right)\right)^n, \quad 1 + \frac{x}{\alpha} \geq n^{-1/\alpha}, \\
\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - d_n}{c_n} \leq x\right) &= \exp\left(-\left(1 + \frac{x}{\alpha}\right)^{-\alpha}\right), \quad 1 + \frac{x}{\alpha} > 0.
\end{aligned}$$

Nähdään, että $F \in \mathrm{MDA}(H_{1/\alpha})$. *Pareto-jakauman rajajakauma on siis Fréchet parametrilla* $\xi = \alpha^{-1}$.

Normeerausvakiot liittyvät yleisesti ottaen alla olevan jakauman kvantiileihin. Vakioiden määrittämisestä ks. [2, luku 3.3].

Lauseet 1.21 ja 1.25 ovat asymptoottisia tuloksia käsitellen maksimien jakaumaa rajalla kun $n \to \infty$, mutta tuottavat välittömästi seuraavan approksimaation:

$$\mathbb{P}(M_n \leq x) = F^n(c_n x + d_n) \approx H_\xi(x), \tag{1.7}$$

suurilla n. Tämä kannustaa seuraavaan lähestymistapaan tilastollisessa mallinnuksessa: kerätään "suuri" määrä riippumattomia havaintoja n:n havainnon $(X_i,\ i = 1, \ldots, n)$ otosmaksimista M_n, ja sovitetaan yleistetty ääriarvojakauma (GEV) maksimihavaintoihin. X_i:t voisivat edustaa esimerkiksi päivittäisiä vedenkorkeusmittauksia, jolloin M_n on vedenkorkeuden vuosimaksimi kun $n = 365$. Tätä kutsutaan ns. blokkimaksimimenetelmäksi (method of block maxima), ja se muodostaa ääriarvoteorian tilastollisten sovellusten klassisen perustan. Menetelmää ja tilastollista estimointia käsitellään osiossa 2.2.

1.3.1 Vaikutuspiirit maksimin suhteen

Useimpien käytännön sovellusten kohdalla riittää todeta, että olennaisesti kaikki (jatkuvat) jakaumat joihin esimerkiksi tilastotieteessä tai vakuutus- ja finanssimatematiikan piirissä törmätään kuuluvat yleistetyn ääriarvojakauman vaikutuspiiriin maksimin suhteen jollekin ξ. Tarkastellaan seuraavaksi kysymystä, mihin kolmesta eri ääriarvojakaumatyypistä eri jakaumat F alla oleville satunnaismuuttujille X johtavat.

Fréchet. Osoittautuu, että Fréchet-jakauman vaikutuspiiriin maksimin suhteen kuuluvat jakaumat voidaan täydellisesti karakterisoida käyttämällä *säännöllisen vaihtelun* (regular variation) ja *hitaan vaihtelun* (slow variation) käsitteitä.

Määritelmä 1.29 (Säännöllisesti ja hitaasti vaihtelevat funktiot)

(i) Funktio $h : (0,\infty) \to (0,\infty)$ *on säännöllisesti vaihteleva indeksillä* $\alpha \in \mathbb{R}$ *(merkitään* $h \in \mathcal{R}_\alpha$*), jos*

$$\lim_{x\to\infty} \frac{h(tx)}{h(x)} = t^\alpha, \quad t > 0.$$

(ii) Funktio $L : (0,\infty) \to (0,\infty)$ *on hitaasti vaihteleva (merkitään* $L \in \mathcal{R}_0$*), jos*

$$\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1, \quad t > 0.$$

Hitaasti vaihtelevat funktiot ovat funktioita, jotka muuttuvat suhteellisen hitaasti suurilla x verrattuna potenssifunktioihin. Esimerkki hitaasti vaihtelevasta funktiosta on logaritmifunktio $\ln(x)$. Säännöllisesti vaihtelevat funktiot voidaan esittää potenssifunktiona kerrottuna hitaasti vaihtelevalla funktiolla, $h(x) = x^\alpha L(x)$.

Otetaan jakauman F hännälle käyttöön merkintä $\bar{F} = 1 - F(x) = \mathbb{P}(X > x)$, $\forall x \in \mathbb{R}$.

Lause 1.30 (Fréchet MDA)
Kaikilla $\xi > 0$*,*

$$F \in \mathrm{MDA}(H_\xi) \iff x_F = \infty \text{ ja } \bar{F} \in \mathcal{R}_{-1/\xi},$$

eli toisin sanoen

$$F \in \mathrm{MDA}(H_\xi) \iff x_F = \infty \text{ ja } \bar{F}(x) = x^{-1/\xi} L(x),$$

jollekin hitaasti vaihtelevalle funktiolle L*.*

Lause 1.30 tarkoittaa mm. että ne jakaumat, jotka johtavat Fréchet-jakaumaan maksimille, ovat jakaumia joilla on säännöllisesti vaihtelevat hännät (negatiivisella indeksillä). Suuretta $\alpha = 1/\xi$ kutsutaan yleisesti jakauman häntäindeksiksi (tail index).

Fréchet MDA:han kuuluvat jakaumat ovat paksuhäntäisiä, ja siten erityisen kiinnostavia useissa vakutuus- ja finanssipuolen sovelluksissa. Olkoon $X \geq 0$ satunnaismuuttuja, jonka kertymäfunktio $F \in \mathrm{MDA}(H_\xi)$, $\xi > 0$. Tällöin $\mathbb{E}(X^k) = \infty$ $\forall k > 1/\xi$. Säännöllisesti vaihteleville hännille $\bar{F} \in \mathcal{R}_{-\alpha}$ nimittäin pätee [2, s. 568]

$$\mathbb{E}\left(X^k\right) < \infty, \quad k < \alpha,$$
$$\mathbb{E}\left(X^k\right) = \infty, \quad k > \alpha,$$

missä siis $X \sim F$. Siis jos jakauma kuuluu $\mathrm{MDA}(H_\xi)$:hin parametrilla $\xi = 1/\alpha > 1$, jakauman ensimmäinen momentti ei ole olemassa (on ääretön); jos $\xi > 1/2$, jakauman varianssi on ääretön, ja jos $\xi > 1/4$, neljäs momentti on ääretön. Fréchet'n vaikutuspiiriin maksimin suhteen kuuluvia jakaumia ovat mm. Fréchet-jakauma itse, Pareto, Cauchy, käänteinen gamma, log-gamma, F ja Burr –jakaumat.

Esimerkki 1.31 (Pareto-jakauma)
Esimerkissä 1.28 osoitettiin suoralla laskulla, että Pareto-jakauma kuuluu Fréchet-jakauman vaikutuspiiriin maksimin suhteen, $F \in \mathrm{MDA}(H_{1/\alpha})$. Pareto-jakauman häntä voidaan saattaa muotoon

$$\bar{F}(x) = \left(\frac{\kappa}{\kappa + x}\right)^{\alpha} = \left(\frac{\kappa + x}{\kappa}\right)^{-\alpha} = x^{-\alpha}\left(\frac{1}{x} + \frac{1}{\kappa}\right)^{-\alpha} = x^{-\alpha}L(x),$$

$\alpha > 0$, $\kappa > 0$, $x \geq 0$. Funktiolle L pätee

$$\lim_{x\to\infty} \frac{L(tx)}{L(x)} = \lim_{x\to\infty} \frac{\left(t^{-1}x^{-1} + \kappa^{-1}\right)^{-\alpha}}{\left(x^{-1} + \kappa^{-1}\right)^{-\alpha}} = \frac{\kappa^{\alpha}}{\kappa^{\alpha}} = 1,$$

eli L on hitaasti vaihteleva ja $\bar{F}$ lauseen 1.30 mukainen.

Weibull. Weibull-jakauman vaikutuspiiriin maksimin suhteen kuuluvilla jakaumilla on äärellinen oikea päätepiste. Seuraava lause karakterisoi vaikutuspiirin.

Lause 1.32 (Weibull MDA)
Kaikilla $\xi < 0$,

$$F \in \mathrm{MDA}(H_{1/\xi}) \Longleftrightarrow x_F < \infty \text{ ja } \bar{F}(x_F - x^{-1}) \in \mathcal{R}_{1/\xi},$$

eli toisin sanoen

$$F \in \mathrm{MDA}(H_{1/\xi}) \Longleftrightarrow x_F < \infty \text{ ja } \bar{F}(x_F - x^{-1}) = x^{1/\xi}L(x),$$

jollekin hitaasti vaihtelevalle funktiolle L.

Weibullin MDA:han kuuluvilla jakaumilla ei ole suurta roolia ääritapahtumien mallinnuksessa vakuutus- ja finassisovelluksissa niiden rajoitetun oikean hännän vuoksi. Vaikka selvästikin tarkasteltavilla suureilla (kuten tappioilla) kaikissa käytännön tilanteissa on olemassa teoreettinen yläraja, kiinnitetyn ylärajan x_F ottamiseen ylimääräiseksi malliparametriksi liittyy ongelmia, ja (oikealta) rajoittamattoman määrittelyalueen omaavat jakaumat tarjoavat enemmän joustavuutta.

Esimerkkejä Weibull-jakauman vaikutuspiiriin maksimin suhteen kuuluvista jakaumista ovat beta-jakauma ja (edellisen erikoistapauksena) tasajakauma.

Esimerkki 1.33 (Tasajakauma)
Tarkastellaan tasajakaumaa, $F \sim \mathrm{T}(0,1)$. Esimerkissä 1.27 todettiin, että $F \in \mathrm{MDA}(H_{-1})$. Välittömästi nähdään, että tasajakaumalle $x_F = 1$ ja

$$\bar{F}(x_F - x^{-1}) = 1 - F(1 - x^{-1}) = x^{-1},$$

eli $L(x) \equiv 1$ ja $\bar{F}(x_F - x^{-1}) \in \mathcal{R}_{-1}$.

Gumbel. Gumbel-jakauman vaikutuspiiriin maksimin suhteen kuuluvien jakaumien karakterisointi osoittautuu vaikeammaksi kuin kahden muun tapauksen. Vaikutuspiiriin kuuluvien jakaumien hännät vaihtelevat suuresti: esimerkiksi normaali- ja lognormaalijakaumat kuuluvat molemmat Gumbelin MDA:han. Normaalijakauma on ohuthäntäinen, mutta lognormaalijakauma jo selvästi paksuhäntäisempi. Hieman epätäsmällisesti Gumbelin vaikutuspiiriin maksimin suhteen kuuluvia jakaumia voitaisiin kuvata jakaumiksi, joiden hännät kuolevat

olennaisesti eksponentiaalista vauhtia. Seuraava lause antaa välttämättömät ja riittävät ehdot.

Lause 1.34 (Gumbel MDA)
$F \in \mathrm{MDA}(H_0)$ jos ja vain jos on olemassa positiivinen funktio $\tilde{a}$ s.e.

$$\lim_{x \nearrow x_F} \frac{\bar{F}\left(x + t\tilde{a}(x)\right)}{\bar{F}(x)} = e^{-t}, \quad t \in \mathbb{R},$$

missä $x_F \leq \infty$.

Eräs mahdollinen valinta funktioksi $\tilde{a}$ on $\tilde{a} = a$, missä

$$a(x) = \int_x^{x_F} \frac{\bar{F}(t)}{\bar{F}(x)} \, \mathrm{d}t, \quad x \leq x_F.$$

Satunnaismuuttujalle $X \sim F$ edellinen on täsmälleen ehdollinen odotusarvo ehdolla $X > x$,

$$a(x) = \mathbb{E}\left(X - x | X > x\right), \quad x < x_F \tag{1.8}$$

Näin määriteltyä funktiota $a(x)$ kutsutaan ylitteen odotusarvofunktioksi (mean excess function), ja sitä tullaan käsittelemään tarkemmin osiossa 2.3 ylitemenetelmän yhteydessä.

Voidaan osoittaa, että satunnaismuuttujalle $X > 0$ jonka jakauma kuuluu Gumbel-jakauman MDA:han pätee $\mathbb{E}(X^k) < \infty$, $\forall k > 0$ (ks. [2, s. 148]). Kaikki positiiviset momentit ovat siis äärellisinä olemassa.

Gumbel-jakauman vaikutuspiiriin maksimin suhteen kuuluvia jakaumia ovat mm. eksponentti, normaali, lognormaali, gamma, χ^2, Gumbel itse ja standardi Weibull-jakauma (erotuksena Weibull-ääriarvojakaumasta).

Yhtenäinen karakterisaatio. Seuraava lause kerää yhteen yleistetyn ääriarvojakauman vaikutuspiiriä maksimin suhteen koskevat tulokset. Sellaisenaan se voidaan nähdä yhtenä klassisen ääriarvoteorian perustuloksista. Lause muodostaa perustan – ja tarjoaa motivoinnin – mm. useille tilastollisille tekniikoille, joista muutamia käsitellään jäljempänä.

Lauseen esittämistä varten tarvitaan kvantiilifunktion käsite. Tätä tullaan käyttämään toistuvasti jatkossa.

Määritelmä 1.35 (Kvantiilifunktio)
Jakauman F yleistettyä käänteisfunktiota,

$$F^{\leftarrow}(t) = \inf\left\{x \in \mathbb{R} | F(x) \geq t\right\}, \quad 0 < t < 1, \tag{1.9}$$

kutsutaan jakauman F kvantiilifunktioksi. Suure $x_t = F^{\leftarrow}(t)$ määrittelee jakauman F t-kvantiilin.

Lause 1.36 (MDA(H_ξ):n karakterisaatio)
Olkoon $\xi \in \mathbb{R}$. Seuraavat väittämät ovat yhtäpitäviä:

(i) $F \in \mathrm{MDA}(H_\xi)$.

(ii) On olemassa positiivinen funktio $a(\cdot)$ siten että, kun $1 + \xi x > 0$,

$$\lim_{u \nearrow x_F} \frac{\bar{F}\left(u + xa(u)\right)}{\bar{F}(u)} = \begin{cases} \left(1 + \xi x\right)^{-\frac{1}{\xi}}, & \xi \neq 0, \\ e^{-x}, & \xi = 0. \end{cases} \tag{1.10}$$

(iii) Kaikilla $x, y > 0$, $y \neq 1$,

$$\lim_{s\to\infty} \frac{U(sx) - U(s)}{U(sy) - U(s)} = \begin{cases} \dfrac{x^\xi - 1}{y^\xi - 1}, & \xi \neq 0, \\ \dfrac{\ln x}{\ln y}, & \xi = 0, \end{cases} \tag{1.11}$$

missä $U(t) = F^{\leftarrow}\left(1 - t^{-1}\right)$, $t > 0$.

Huomautus 1.37

1) Kohta (iii) karakterisoi GEV-jakauman vaikutuspiirin maksimin suhteen funktion U asymptoottisen käytöksen kautta. Muotoilemalla esitys (1.11) *uudelleen päädytään menetelmään datan ulkopuolisten kvantiilien estimoimiseksi (ks. [2, s. 349-350]).* (1.11)*:n erikoistapaus toimii myös motivaationa muotoparametrin* ξ *ns. Pickands-estimaattorille (ks. osio 3.2).*

2) Kun $x > 0$, *yhtälöllä* (1.10) *on kiintoisa tulkinta todennäköisyyksien kautta: Olkoon X satunnaismuuttuja jakaumalla* $F \in \mathrm{MDA}(H_\xi)$. *Tällöin* (1.10) *voidaan yhtäpitävästi kirjoittaa*

$$\lim_{u \nearrow x_F} \mathbb{P}\left(\frac{X-u}{a(u)} > x | X > u\right) = \begin{cases} (1+\xi x)^{-\frac{1}{\xi}}, & \xi \neq 0, \\ e^{-x}, & \xi = 0. \end{cases} \tag{1.12}$$

Siten (1.10) *antaa tason u skaalattujen ylitteiden jakauman, kun taso u kasvaa äärettömäksi — tai käytännön sovelluksia ajatellen approksimaation skaalattujen ylitteiden jakaumalle kun kynnys u on korkea. Skaalaustekijänä toimii kynnystasosta riippuva* $a(u)$. (1.10) *motivoi ns. yleistetyn Pareto-jakauman, jolla on aivan keskeinen asema modernimmassa ääriarvoteoriassa korkean tason ylittävien ilmiöiden ylitteiden suuruusjakauman mallina. Esimerkkinä voidaan ajatella vaikkapa omapidätysrajan ylittävien tappioiden jakaumaa Excess of loss -jälleenvakuutuksessa.*[7] *Yleistettyä Pareto-jakaumaa käsitellään osiossa 1.5.*

1.4 Stationaariset prosessit

Klassinen ääriarvoteoria koskee riippumattomia ja samoin jakautuneita muuttujia, ja edellä esitetty tarkastelu perustuu iid-oletukselle. Käytännön sovelluksissa riippumattomuusoletus ei kuitenkaan usein ole perusteltavissa. Ajallinen riippuvuus on yleistä havaintoaikasarjojen maksimeissa autokorrelaation, muiden muuttujien vaikutuksen tai molempien seurauksena. Lyhyen kantaman riippuvuus johtaa siihen, että äärihavainnot pyrkivät kasaantumaan ajallisesti: esimerkiksi sijoitusinstrumenttien tuottoaikasarjat ilmentävät usein volatiliteetin

[7] Excess of loss (XL) -jälleenvakuutuksessa (suomalaisittain joskus *yksittäisylivahinkojälleenvakuutus*) kukin yksittäinen jälleenvakuutuksen piiriin kuuluva vahinko jaetaan *ensivakuuttajan* eli alkuperäisen vakuuttajan ja *jälleenvakuuttajan* kesken. Olkoon jälleenvakuutussopimuksessa määritelty omavastuuraja $M > 0$ – tästä käytetään vakiintuneesti termejä excess- eli xs-point, attachment point tai priority. Merkitään yksittäistä vahinkoa X: XL-järjestelyssä ensivakuuttajan vastuulle jää tästä määrä $X^{ov} = \min(X, M)$, ja jälleenvakuuttajalle $X^{jv} = \max(X - X^{ov}, 0)$ eli mahdollinen omavastuurajan ylittävä osuus. (Käytännössä jälleenvakuuttajan vastuulle asetetaan yleensä myös yläraja (limit tai exit point), merk. A – tällöin XL-sopimusta nimitetään "M xs A".)

kasaantumista (volatility clustering), siten että volatiliteetin ollessa "korkealla" havaitaan todennäköisemmin itseisarvoltaan suuria hintamuutoksia lyhyellä aikavälillä kuin volatiliteetin ollessa "matalalla".[8] Samoin esimerkiksi joen virtausmaksimit sattuvat yleensä lähekkäin myrskyn jälkeen.

Riippumattomien ja samoin jakautuneiden satunnaismuuttujien jonon luonnollinen yleistys on tarkastella (vahvasti) stationaarista prosessia. Satunnaismuuttujajonon $(X_t)_{t\in\mathbb{Z}}$ sanotaan olevan *vahvasti stationaarinen*, jos sen äärellisulotteiset jakaumat ovat muuttumattomia ajassa siirtymisen suhteen, eli

$$(X_{t_1},\ldots,X_{t_n}) \stackrel{d}{=} (X_{t_1+h},\ldots,X_{t_n+h})$$

kaikilla indeksivalinnoilla $t_1 < \cdots < t_n$, $\forall h \in \mathbb{Z}$, $\forall n \in \mathbb{N}$. Sovelluksia ajatellen jonoa (X_t) voidaan ajatella tasavälisten havaintojen muodostamana aikasarjana, jossa h-pituisen havaintoblokin $(X_t, X_{t+1}, \ldots, X_{t+h})$ jakauma on sama ajanhetkestä riippumatta eli kaikille $t \in \mathbb{Z}$.

Satunnaismuuttujajono on *heikosti stationaarinen*, jos

$$\begin{aligned} \mathbb{E}(X_t) &= \mu, && \forall t \in \mathbb{Z}, \\ \operatorname{Cov}(X_t, X_s) &= \operatorname{Cov}(X_{t+k}, X_{s+k}), && \forall t,s,k \in \mathbb{Z}, \end{aligned}$$

olettaen että kaksi ensimmäistä momenttia ovat olemassa. Vahvasti stationaarinen prosessi on myös heikosti stationaarinen. Jatkossa tarkastellaan vahvasti stationaarista jonoa, ja käytetään tästä lyhyesti nimitystä "stationaarinen".

Stationaarisen prosessin muodostavat satunnaismuuttujat voivat siis ilmentää keskinäistä riippuvuutta, mutta tällaisen prosessin tilastolliset ominaisuudet pysyvät samanlaisina (edellä esitettyjen määritelmien mielessä) ajan suhteen. Stationaarisuus on selvästi riippumattomuutta realistisempi oletus monia fysikaalisia prosesseja ajatellen. Osoittautuu lisäksi, että tiettyjen lisäehtojen pätiessä stationaaristen jonojen kohdalla päädytään saman tyyppisiin rajajakaumiin kuin iid tapauksessakin. Tilastollisen mallinnuksen kannalta tulos tulee tarkoittamaan, että stationaariseen tapaukseeen voidaan soveltaa pitkälti samoja menetelmiä kuin iid tapaukseenkin.

Olkoon (X_n) stationaarinen prosessi jossa satunnaismuuttujien X_i reunajakauma on F, ja $M_n = \max(X_1,\ldots,X_n)$ niin kuin aiemminkin. Merkitään $(\tilde{X}_n)$:llä vastaavaa riippumattomien ja samoin jakautuneiden satunnaismuuttujien muodostamaa prosessia samalla reunajakaumalla F; sanotaan, että $(\tilde{X}_n)$ on (X_n):ään liittyvä iid-prosessi. Tämän otosmaksimia merkitään vastaavasti $\tilde{M}_n = \max(\tilde{X}_1,\ldots,\tilde{X}_n)$.

Riippuvuus stationaarisessa jonossa voi ilmetä monissa eri muodoissa, ja on mahdotonta rakentaa yleistä äärimmäisten arvojen teoriaa kaikkien stationaaristen jonojen muodostamalle luokalle asettamatta joitain rajoitteita mahdollisille riippuvuusrakenteille (ks. [2, s. 210-211]). On luonnollista kysyä, millä ehdoilla jonon (X_n) maksimien rajakäyttäytyminen on identtistä vastaavan iid jonon $(\tilde{X}_n)$ maksimien kanssa. Leadbetter et. al osoittivat, että tähän tarvitaan kaksi alla tarkemmin määriteltyä teknistä ehtoa (ks. [4]): Jos stationaarinen jono ilmentää vain heikkoa pitkän kantaman riippuvuutta (long-range dependence)

[8] Ks. tarkemmin luku 4.

äärimmäisillä tasoilla (ehdon $D(u_n)$ mielessä), ja jos se ei osoita taipumusta korkeiden arvojen kasaantumiseen (ehdon $D'(u_n)$ mielessä), niin jonojen (X_n) ja $(\tilde{X}_n)$ maksimeilla on identtinen rajajaukauma.

Esitetään ehtojen formaalit määritelmät.

Määritelmä 1.38 (Ehto $D(u_n)$; distributional mixing condition) *Sanotaan, että jono (X_n) täyttää ehdon $D(u_n)$, jos kaikille $p, q, n \in \mathbb{N}$ ja kaikille*

$$1 \leq i_1 < \cdots < i_p < j_1 < \cdots < j_q \leq n$$

siten, että $j_1 - i_p > l$, pätee

$$\left| \mathbb{P}\left(\max_{i \in A_1 \cup A_2} X_i \leq u_n \right) - \mathbb{P}\left(\max_{i \in A_1} X_i \leq u_n \right) \mathbb{P}\left(\max_{i \in A_2} X_i \leq u_n \right) \right| \leq \alpha_{n,l},$$

missä $A_1 = \{i_1, \ldots, i_p\}$, $A_2 = \{j_1, \ldots, j_q\}$ ja $\alpha_{n,l} \to 0$ kun $n \to \infty$ jollekin jonolle l_n, jolle pätee $l_n/n \to 0$, $n \to \infty$.

Ehto $D(u_n)$ tarkoittaa, että jonon (X_n) muodostavat satunnaismuuttujat ilmentävät tietynlaista asymptoottista riippumattomuutta. Riippumattomien satunnaismuuttujien jonolle todennäköisyyksien erotus määritelmässä 1.38 on täsmälleen nolla mille tahansa kynnystasojen jonolle (u_n). Yleisemmin vaaditaan vain, että ehto $D(u_n)$ pätee tietylle n:n suhteen kasvavalle jonolle (u_n). Tällaiselle kynnysjonol-le ehto intuitiivisesti varmistaa, että riittävän etäällä toisistaan oleville satunnaismuuttujajoukoille todennäköisyyksien ero yllä on niin pieni, ettei sillä ole vaikutusta ääriarvojen rajajakaumaan.

Määritelmä 1.39 (Ehto $D'(u_n)$; anti-clustering condition) *Sanotaan, että ehto $D'(u_n)$ pätee jonolle (X_n), jos*

$$\lim_{k \to \infty} \limsup_{n \to \infty} n \sum_{j=2}^{\lfloor n/k \rfloor} \mathbb{P}\left(X_1 > u_n, X_j > u_n\right) = 0$$

Yllä $\lfloor \cdot \rfloor$ tarkoittaa kokonaislukuosaa. Ehtoa $D'(u_n)$ kutsutaan "kasaantumisenestoehdoksi" stationaariselle jonolle (X_n), koska ehdosta seuraa, että (korkean) tason u_n yhteiset ylitykset parille (X_i, X_j) tulevat odotusarvoisesti hyvin epätodennäköisiksi suurille n.[9]

Tarkastellaan ehtojen suhdetta käytännön sovelluksiin: Ehto $D(u_n)$ rajoittaa pitkän kantaman riippuvuutta äärimmäisillä tasoilla, siten että tapahtumat $(X_i > u)$ ja $(X_j > u)$ ovat lähes riippumattomia riittävän suurella u, kun ajanhetket i ja j ovat riittävän kaukana toisistaan eli $|i - j|$ on suuri. Monet stationaariset jonot täyttävät tämän ehdon. Lisäksi sen voidaan usein katsoa olevan perusteltu fysikaalisten prosessien tapauksessa: esimerkiksi rankkasateen sattuminen tänään voi lisätä äärimmäisen kovan sateen sattumistodennäköisyyttä päivän tai parin sisällä, mutta tuskin tiettynä päivänä vuoden päästä. Yleisesti ottaen äärimmäisten ilmiöiden pitkän kantaman riippuvuus vaikuttaa epätodennäköiseltä useimmissa tapauksissa, geneettistä dataa mahdollisesti lukuun ottamatta [14].

[9] Ehto implikoi, että $\mathbb{E}\sum_{1 \leq i < j \leq \lfloor n/k \rfloor} \mathbb{1}(X_i > u_n, X_j > u_n) \leq \lfloor n/k \rfloor \sum_{j=2}^{\lfloor n/k \rfloor} \mathbb{E}\mathbb{1}(X_i > u_n, X_j > u_n) = \lfloor n/k \rfloor \sum_{j=2}^{\lfloor n/k \rfloor} \mathbb{P}\left(X_i > u_n, X_j > u_n\right) \to 0$; ks. [2, s. 213].

Pitkän kantaman riippuvuuden poissulkeminen keskittää tarkastelun lyhyen kantaman riippuvuuteen (short-range dependence) ääriarvoissa. Tämä tyypillisesti ilmenee siten, että suuret arvot pyrkivät kasaantumaan, ts. sattuvat klustereissa. Jos ehto $D'(u_n)$ pätee, stationaarisella prosessilla ei ole taipumusta suurien arvojen kasaantumiseen. Ehdot $D(u_n)$ ja $D'(u_n)$ yhdessä varmistavat, että ehdot täyttävän stationaarisen prosessin ääriarvoilla on sama asymptoottinen käyttäytyminen kuin vastaavalla iid jonolla. Tällöin stationaariseen prosessiin pätee siis täsmälleen sama teoria kuin vastaavaan iid satunnaismuuttujajonoonkin.

Käytännössä kuitenkin ehto $D'(u_n)$ ei usein täyty reaalimaailman prosesseja tarkastellessa. Näin on esimerkiksi useimpien finanssiaikasarjojen kohdalla, joissa usein havaitaan esimerkiksi volatiliteetin kasautumisesta johtuvaa suurien havaintoarvojen ajallista kasaantumista. Esimerkiksi stationaariset ARCH-prosessit[10] täyttävät ehdon $D(u_n)$, mutteivät ehtoa $D'(u_n)$. Äärihavaintojen kasaantuminen lyhyen kantaman riippuvuuden seurauksena (ehdon $D'(u_n)$ täyttymättä jäämisen mielessä) vaatii lisäkäsittelyä jotta teoriaa voidaan soveltaa. Ns. ääriarvoindeksi (extremal index) osoittautuu avainkäsitteeksi. Tarkkaa määritelmää varten esitetään ensin vaihtoehtoinen ehto maksimien suppenemiselle iid tapauksessa: Ottamalla logaritmi lausekkeessa (1.6) (ja käyttämällä lauseen 1.25 muotoa), saadaan

$$\begin{aligned} \big(1-\bar{F}(c_n x+d_n)\big)^n &\xrightarrow{d} H_\xi(x), \\ n\ln\big(1-\bar{F}(c_n x+d_n)\big) &\xrightarrow{d} \ln H_\xi(x), \end{aligned} \tag{1.13}$$

ja edelleen, käyttämällä relaatiota $\ln(1-x)\sim -x$, $x\to 0$, saadaan vaihtoehtoiseksi ehdoksi vakioiden $c_n>0$ ja $d_n\in\mathbb{R}$ olemassaolo s.e.

$$\lim n\bar{F}(c_n x+d_n) = -\ln H_\xi(x), \quad x\in\mathbb{R}. \tag{1.14}$$

(Kun $H_\xi(x)=0$, raja-arvoksi tulkitaan ∞.) Itse asiassa pätee seuraava yleisemmässä muodossa oleva tulos:

Propositio 1.40 (Poisson-approksimaatio)
Annetulle $\tau\in[0,\infty]$ ja ei-vähenevälle jonolle (u_n), $u_n\in\mathbb{R}$, seuraavat ovat yhtäpitäviä:

$$\lim_{n\to\infty} n\bar{F}(u_n) = \tau, \tag{1.15}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\tilde{M}_n\le u_n\right) = e^{-\tau}. \tag{1.16}$$

Seuraava tarkastelu tarjoaa taustaa propositiolle: Oletetaan yksinkertaisuuden vuoksi, että $0<\tau<\infty$ ja määritellään $B_n=\sum_{i=1}^n \mathbb{1}(\tilde{X}_i>u_n)$. Yksittäinen ylite on Bernoulli-jakautunut, ja tason u_n ylitteiden lukumäärällä B_n on binomijakauma parametreilla $(n,\bar{F}(u_n))$. Standardi Poisson-jakaumaa binomijakauman rajana koskeva tulos[11] sanoo nyt, että $B_n\to \mathrm{Poi}(\tau)$ jos ja vain jos $\mathbb{E}(B_n)=n\bar{F}(u_n)\to\tau$. Tämä on juuri (1.15). Edelleen saadaan $\mathbb{P}(\tilde{M}_n\le u_n)=\mathbb{P}(B_n=0)\to e^{-\tau}$. Esitetyn vuoksi (1.16):tä kutsutaan joskus Poisson-approksimaatioksi todennäköisyydelle $\mathbb{P}(\tilde{M}_n\le u_n)$. [2, s. 116]

[10] ARCH = AutoRegressive Conditional Heteroskedasticity; ks. kappale 4.3.1
[11] Ks. esim. [15, Lemma 3.1.1.].

Proposition 1.40 yhtäpitävät väitteet iid-tapauksessa tarjoavat suoran vertailukohdan stationaariseen tapaukseen nähden.

Määritelmä 1.41 (Ääriarvoindeksi, extremal index)
Olkoon (X_n) aidosti stationaarinen jono ja $0 \leq \theta \leq 1$. Oletetaan, että kaikille $\tau > 0$ on olemassa jono $(u_n) = (u_n(\tau))$ siten, että

$$\lim_{n\to\infty} n\bar{F}(u_n) = \tau, \tag{1.17}$$

$$\lim_{n\to\infty} \mathbb{P}(M_n \leq u_n) = e^{-\theta\tau}. \tag{1.18}$$

Tällöin sanotaan, että θ on jonon (X_n) ääriarvoindeksi.

Huomautus 1.42 *Voidaan osoittaa, että määritelmän 1.41 mukainen θ on hyvin määritelty, eikä riipu jonon $(u_n(\tau))$ valinnasta. [2, s. 416-417]*

Vertaamalla propositiota 1.40 iid-tapaukselle ja määritelmää 1.41 stationaariselle jonolle ääriarvoindeksillä θ, nähdään, että

$$\mathbb{P}(M_n \leq u_n) \approx \mathbb{P}^\theta(\tilde{M}_n \leq u_n) = F^{\theta n}(u_n), \tag{1.19}$$

olettaen että $n\bar{F}(u_n) \to \tau > 0$. Tämä voidaan nähdä ääriarvoindeksin heuristisena määritelmänä: (1.19):n mukaan n:n havainnon maksimi stationaarisesta prosessista ääriarvoindeksillä θ vastaa $n\theta(< n)$ havainnon maksimia vastaavasta iid jonosta. Luvun $n\theta$ voi siis tässä mielessä ajatella ilmoittavan "riippumattomien" havaintoryppäiden (klusterien) lukumäärän n havainnon otoksessa. Tällöin indeksillä θ on tulkinta klusterin keskimääräisen koon käänteislukuna.

Aidosti stationaarisia jonoja koskeva päätulos on seuraava:

Lause 1.43 (Stationaarisen jonon rajajakauma)
Jos (X_n) on aidosti stationaarinen jono ääriarvoindeksillä $\theta > 0$ ja $(\tilde{X}_n)$ vastaava iid jono, niin

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\tilde{M}_n - d_n}{c_n} \leq u_n\right) = H(x) \Longleftrightarrow \lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - d_n}{c_n} \leq u_n\right) = H^\theta(x) \tag{1.20}$$

ei-degeneroituneille jakaumille $H(x)$, $H^\theta(x)$.

Jos siis $F \in \mathrm{MDA}(H_\xi)$, niin stationaarisen prosessin normeerattujen maksimien asymptoottinen jakauma on myös yleistetty ääriarvojakauma. Muotoparametri ξ säilyy samana kuin iid-tapauksessakin, mutta lokaatio- ja skaalaparametrit muuttuvat: kun $\xi \neq 0$,

$$\begin{aligned}
H_\xi^\theta(x) &= \left(\exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\}\right)^\theta \\
&= \exp\left\{-\theta\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\} \\
&= \exp\left\{-\left(1+\theta^{-\xi}+\theta^{-\xi}\xi\frac{x-\mu}{\sigma}-1\right)^{-1/\xi}\right\} \\
&= \exp\left\{-\left(1+\xi\frac{x-\tilde{\mu}}{\tilde{\sigma}}\right)^{-1/\xi}\right\},
\end{aligned}$$

missä $\tilde{\mu} = \mu - \frac{\sigma}{\xi}(1-\theta^{\xi})$ ja $\tilde{\sigma} = \sigma\theta^{\xi}$.[12]

Kaikilla aidosti stationaarisilla prosesseilla ei ole olemassa ääriarvoindeksiä, mutta yleensä näin on. Ks. [2, s. 418]. Esitetään joitain esimerkkejä ([5, s. 271]):

- Riippumattomilla ja samoin jakautuneilla satunnaismuuttujajonoilla $\theta = 1$.
- ARMA-prosesseilla normaalijakautuneilla innovaatioilla $\theta = 1$; jos kuitenkin innovaatioiden jakauma kuuluu Fréchet MDA:han, niin $\theta < 1$.
- ARCH- ja GARCH-prosesseilla $\theta < 1$.

Mallinnusta ajatellen tarvitsee olennaisesti ottaen erottaa vain tapaukset $\theta = 1$ ja $\theta < 1$: edellisessä aikasarjalla ei ole taipumusta kasaantua suurilla arvoilla ja maksimit käyttäytyvät täsmälleen kuten iid-tapauksessakin; jälkimmäisessä klusteroituminen vaikuttaa rajajakauman parametrien arvoihin ääriarvoindeksin kautta. Käytännössä tällä ei ole merkitystä, koska parametrit täytyy joka tapauksessa estimoida datasta. Satunnaismuuttujien X_i riippuvuus kuitenkin aiheuttaa sen, että konvergenssi GEV-jakaumaan on hitaampaa, koska efektiivisesti (riippumaton) otoskoko on n:n sijasta vain $n\theta$. Siten stationaarisen prosessin kohdalla tarvitaan yleisesti enemmän dataa kuin iid prosessin tapauksessa. Estimointia varten muodostettujen havaintoblokkien tulisi myös olla riittävän pitkiä, jotta eri blokkien maksimeja voidaan pitää olennaisesti riippumattomina havaintoina tilastollisia menetelmiä (suurimman uskottavuuden menetelmää) ajatellen.

1.5 Ylitteet ja yleistetty Pareto-jakauma

Edellä tarkasteltiin yleistettyä ääriarvojakaumaa (GEV) normalisoitujen otosmaksimien asymptoottisena jakaumana. Tilastollisena sovelluksena tämä tarkoittaa datan jakamista n:n havainnon blokkeihin, ja kunkin blokin maksimihavainnon poimimista analyysia varten. Tämän lähestymistavan yhtenä ongelmana on kuitenkin se, että paljon dataa "heitetään hukkaan": esimerkiksi tiettynä vuotena saattaa sattua useita havaintoja, jotka ovat toisen vuoden maksimia korkeampia, mutta jäävät menetelmässä huomiotta. Vaihtoehtoinen lähestymistapa ääriarvoihin on tarkastella kaikkia tietyn korkean rajan u ylittäviä havaintoja. Tätä kutsutaan ylitemenetelmäksi (method of treshold excesses), ja sen tilastollista toteutusta käsitellään kappaleessa 2.3.

Relaation (1.10) oikea puoli motivoi seuraavan määritelmän:

Määritelmä 1.44 (Yleistetty Pareto-jakauma, GPD)
Määritellään jakauma G_ξ ehdosta

$$G_\xi(x) = \begin{cases} 1-(1+\xi x)^{-1/\xi}, & \xi \neq 0, \\ 1-e^{-x}, & \xi = 0. \end{cases}$$

[12] Kun $\xi = 0$, saadaan vastaavasti $H_0^\theta(x) = \left(\exp\left\{e^{-\frac{x-\mu}{\sigma}}\right\}\right)^\theta = \exp\left\{\theta e^{-\frac{x-\mu}{\sigma}}\right\} = \exp\left\{e^{-\frac{x-\tilde{\mu}}{\sigma}}\right\}$, missä $\tilde{\mu} = \mu + \sigma\ln\theta$.

G_ξ on nimeltään standardi yleistetty Pareto-jakauma (generalized Pareto distribution, GPD), missä ξ on jakauman muotoparametri. Kun lisäksi otetaan käyttöön skaalaparametri $\beta > 0$, saadaan yleistetty Pareto-jakauma $G_{\xi,\beta}$,

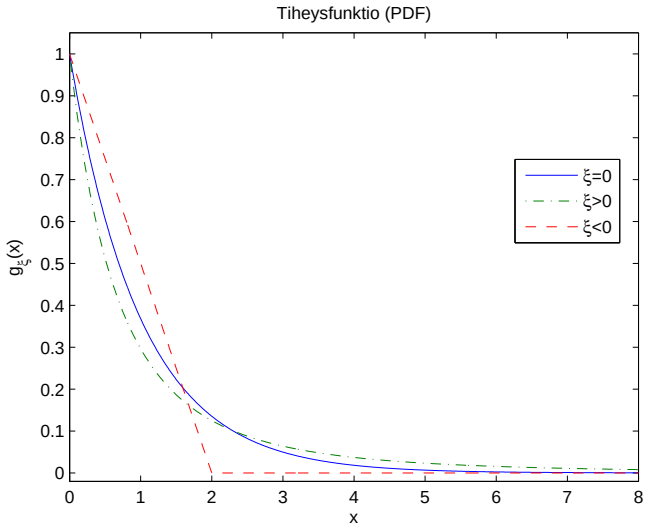
$$G_{\xi,\beta}(x) = \begin{cases} 1 - \left(1 + \xi\frac{x}{\beta}\right)^{-\frac{1}{\xi}}, & \xi \neq 0, \\ 1 - e^{-x/\beta}, & \xi = 0. \end{cases}$$

$G_{\xi,\beta}$:n määrittelyalue on

$$\begin{aligned} x > 0, &\quad \text{kun} \quad \xi \geq 0, \\ 0 \leq x \leq -\tfrac{\beta}{\xi}, &\quad \text{kun} \quad \xi < 0. \end{aligned}$$

Jakaumaperhe $G_{\xi,\beta}$ voidaan laajentaa ottamalla mukaan lokaatioparametri; olkoon tämä $\nu \in \mathbb{R}$. Tällöin määritellään $G_{\xi,\nu,\beta}(x) = G_{\xi,\beta}(x - \nu)$.

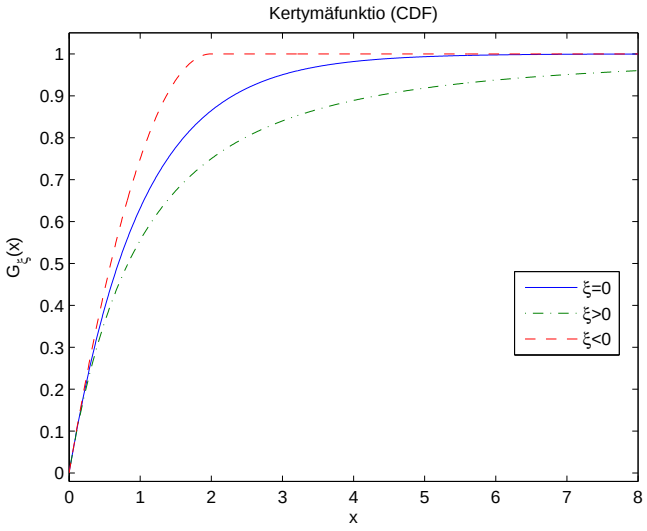
Yleistetyn Pareto-jakauman parametrisointi on ξ:n suhteen jatkuva kiinteällä x, eli $\lim_{\xi\to 0} G_{\xi,\beta}(x) = G_{0,\beta}(x)$. Tämä on erityisen hyödyllistä tilastollisia menetelmiä ja suurimman uskottavuuden estimointia ajatellen. Kuten GEV, GPD on "yleistetty" siinä mielessä, että sen parametrisointi sisältää kolme eri erikoistapausta: kun $\xi > 0$, $G_{\xi,\beta}$ on tavallinen Pareto-jakauma parametreilla $\alpha = 1/\xi$ ja $\kappa = \beta/\xi$; kun $\xi = 0$, $G_{\xi,\beta}$ on eksponenttijakauma parametrilla $1/\beta$; ja kun $\xi < 0$, $G_{\xi,\beta}$ on lyhythäntäinen ($x_F = -\beta/\xi$) tyypin II Pareto-jakauma. Kuvassa 1.3 on esitetty GP-jakauman G_ξ tiheysfunktio ja kuvassa 1.4 vastaavasti kertymäfunktio eri ξ:n arvoja vastaten.



Kuva 1.3: GP-jakauman tiheysfunktio.

Yleistetylle Pareto-jakaumalle pätee $G_{\xi,\beta} \in \mathrm{MDA}(H_\xi)$ kaikilla $\xi \in \mathbb{R}$; tapauksissa $\xi > 0$ ja $\xi < 0$ tämä seuraa suoraan lauseista 1.30 ja 1.32. Kun $\xi > 0$, voidaan näyttää että

$$\mathbb{E}\left(X^k\right) = \infty, \quad \forall k \geq 1/\xi > 0,$$



Kuva 1.4: GP-jakauman kertymäfunktio.

missä $X \sim G_{\xi,\beta}$. Yleistetyn Pareto-jakauman odotusarvo on siis olemassa, mikäli $\xi < 1$, ja on tällöin

$$\mathbb{E}(X) = \frac{\beta}{1-\xi}. \tag{1.21}$$

Yleistetty Pareto-jakauma toimii ääriarvoteoriassa luonnollisena asymptoottisena mallina korkean kynnyksen ylittävien arvojen jakaumalle (ylitejakaumalle). Määritellään jälkimmäinen muodollisesti.

Määritelmä 1.45 (Ylitejakauma)
Olkoon X satunnaismuuttuja jakaumalla F. Kynnyksen u ylitteiden jakauma (excess distribution over the threshold u), tai ylitejakauma, on

$$F_u(x) = \mathbb{P}(X - u \leq x | X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}, \quad 0 \leq x < x_F - u, \tag{1.22}$$

missä $x_F \leq \infty$ on jakauman F oikea päätepiste.

Vakuutuskontekstissa F_u tunnetaan yleisesti excess-of-loss –jakaumana ("vahinkoylitejaukama"). Ylitejakaumaan läheisesti liittyvä käsite on ylitteen odotusarvofunktio (mean excess function), joka nimensä mukaisesti on annetun tason ylitteen suuruuden ehdollinen odotusarvo, ehdolla että taso ylitetään. Tähän törmättiin jo osiossa 1.3.1; ks. yhtälö (1.8). Esitetään tämä vielä määritelmänä.

Määritelmä 1.46 (Ylitteen odotusarvofunktio)
Satunnaismuuttujan X ylitteen odotusarvofunktio *on*

$$e(u) = \mathbb{E}(X - u | X > u), \tag{1.23}$$

mikäli odotusarvo on olemassa.

Määritetään ylitejakauma ja ylitteen odotusarvofunktio yleistetylle Pareto-jakaumalle. Olkoon X satunnaismuuttuja jakaumalla $F = G_{\xi,\beta}$. Suoralla laskulla saadaan

$$\begin{aligned} F_u(x) &= \frac{G_{\xi,\beta}(x+u) - G_{\xi,\beta}(u)}{1 - G_{\xi,\beta}(u)} = \frac{\left(1+\xi\frac{u}{\beta}\right)^{-1/\xi} - \left(1+\xi\frac{x+u}{\beta}\right)^{-1/\xi}}{\left(1+\xi\frac{u}{\beta}\right)^{-1/\xi}} \\ &= 1 - \left(\frac{1+\xi(x+u)/\beta}{1+\xi u/\beta}\right)^{-1/\xi} = 1 - \left(1 + \frac{\xi x/\beta}{1+\xi u/\beta}\right)^{-1/\xi} \\ &= 1 - \left(1+\xi\frac{x}{\beta+\xi u}\right)^{-1/\xi}, \end{aligned}$$

missä $0 \le x < \infty$ kun $\xi > 0$ ja $0 \le x \le -\beta/\xi - u$ kun $\xi < 0$.[13] Siis $F_u(x) = G_{\xi,\beta(u)}(x)$, missä $\beta(u) = \beta + \xi u$. Ylitejakauma on siis myös GPD, samalla muotoparametrilla ξ, mutta skaalauksella joka kasvaa lineaarisesti kynnystason u suhteen. Samanlainen tulos pätee myös asetettaessa (korkea) kynnys u perustasoksi, ja tarkasteltaessa minkä tahansa tätä korkeamman kynnyksen $v > u$ ylityksiä. Tarkastellaan mukavuuden vuoksi jakauman häntiä:

$$\begin{aligned} \bar{F}_v(x) &= \frac{\bar{F}(v+x)}{\bar{F}(v)} = \frac{\bar{F}(u+(v+x-u))}{\bar{F}(u)}\frac{\bar{F}(u)}{\bar{F}(u+(v-u))} = \frac{\bar{F}_u(v+x-u)}{\bar{F}_u(v-u)} \\ &= \frac{\bar{G}_{\xi,\beta}(v-u+x)}{\bar{G}_{\xi,\beta}(v-u)} = \bar{G}_{\xi,\beta+\xi(v-u)}(x). \end{aligned}$$

Siis $F_v(x) = G_{\xi,\beta(v-u)}(x)$, missä $\beta(v-u) = \beta+\xi(v-u)$, kun $v > u$ (nyt siis skaalaparametri β vastaa tasoa u, jota korkeampia ylitteitä tarkastellaan).

GP-jakauman ylitteen odotusarvofunktio on vastaavasti

$$e(u) = \frac{\beta(u)}{1-\xi} = \frac{\beta+\xi u}{1-\xi},$$

missä missä $0 \le u < \infty$, kun $0 \le \xi < 1$ ja $0 \le u \le -\beta/\xi$, kun $\xi < 0$. Ylitteen odostusarvofunktio $e(u)$ on siis kynnyksen u lineaarinen funktio. Lineaarisuus säilyy tarkastellessa tasoa u korkeampia kynnystasoja $v > u$:

$$e(v) = \frac{\beta+\xi(v-u)}{1-\xi} = \frac{\xi v}{1-\xi} + \frac{\beta-\xi u}{1-\xi}, \tag{1.24}$$

missä nyt $u \le v < \infty$, kun $0 \le \xi < 1$ ja $u \le v \le u - \beta/\xi$, kun $\xi < 0$. Tämä on yleistetyn Pareto-jakauman karakteristinen ominaisuus, ja $e(u)$:n empiiriseen vastineeseen perustuvaa graafista tekniikkaa tullaan hyödyntämään kynnyksen u valinnassa (etsimällä datan kuvaajasta alue, josta alkaen funktio on likimain lineaarinen) luvussa 2.3.

Edellinen tarkastelu osoittaa, että yleistetyllä Pareto-jakaumalla on tietynlainen stabiilius-ominaisuus ylitejakauman suhteen. Osoittautuu, että GPD on luonnollinen ylitteiden rajajakauma monille alla oleville jakaumille F. Seuraava perustulos voidaan nähdä MDA(H_ξ):n karakterisaationa kaikille $\xi \in \mathbb{R}$ ylitteiden jakaumien asymptoottisen käyttäytymisen avulla.

[13] Kun $\xi = 0$, päädytään eksponenttijakaumaan, ja tunnetusti $F_u(x) = F(x)$, $\forall x$ (eksponenttijakaumalla "ei ole muistia").

Lause 1.47 (Pickands–Balkema–de Haan)
On olemassa positiivinen funktio $\beta(u)$ siten, että

$$\lim_{u \to x_F} \sup_{0 \le x < x_F - u} \left| F_u(x) - G_{\xi,\beta(u)}(x) \right| = 0,$$

jos ja vain jos $F \in \mathrm{MDA}(H_\xi)$, $\xi \in \mathbb{R}$.

Lause 1.47 sanoo, että jakaumat, joille normeeratut maksimit suppenevat yleistettyyn ääriarvojakaumaan, ovat jakaumia, joille ylitteiden jakauma suppenee yleistettyyn Pareto-jakaumaan kun kynnystä kasvatetaan riittävästi. Ylitteiden asymptoottisen GP-jakauman muotoparametri ξ on lisäksi sama kuin maksimien asymptoottisen GEV-jakauman vastaava. Lause osoittautuu laajasti käyttökelpoiseksi, ja sen perusteella yleistettyä Pareto-jakaumaa voidaan pitää korkean kynnyksen ylittävien arvojen jakauman perustavanlaatuisena mallina.

Tilastollisia sovelluksia ajatellen lause 1.47 antaa välittömästi seuraavan approksimaation ylitejakaumalle F_u, suurilla u:

$$F_u(x) = \mathbb{P}(X - u \le x | X > u) \approx G_{\xi,\beta(u)}(x), \quad x > 0, \tag{1.25}$$

missä $\beta(u)$ – ja ξ – on estimoitava datasta. Vaihtoehtoisesti voidaan tarkastella koko ylityksien suuruutta ($u + (x - u) = x$) eikä vain ylitteitä ($x - u$): kun $x > u$,

$$\mathbb{P}(X > x | X > u) \approx \bar{G}_{\xi,u,\beta(u)}(x).$$

Ylitteisiin perustuvaa tilastollista mallinnusta tarkastellaan luvussa 2.3.

1.6 Pisteprosessimallit

Edellisessä osiossa tarkasteltiin vain annetun, korkean tason ylitteiden suuruuksia, ja esitettiin yleistetty Pareto-jakauma laajasti soveltuvana asymptoottisena jakaumana näille. Tarkastelua on kuitenkin luonnollista laajentaa koskemaan paitsi ylitteiden suuruuksia, myös niiden sattumista ajassa. Ajatellaan vaikkapa jälleenvakuuttajan maksettavaksi tulevien, ensivakuuttajan omapidätysrajan ylittävien vahinkojen sattumista XL-jälleenvakuutuksessa.[14] Korkean kynnyksen ylittävien arvojen sattumista ajassa on luonnollista kuvata pisteprosesseja käyttäen. Osoittautuu, että tämä lähestymistapa yleistää edellä käsitellyt mallit maksimeille ja ylitteille, tarjoten yhtenäisen mallinnuskehyksen joka sulkee sisäänsä sekä GEV- että GPD-mallit.

Aikaulottuvuus oli itse asiassa implisiittisesti mukana myös edellisessä osiossa ylitteiden jakaumaa tarkastellessa: proposition 1.40 mukaan korkean tason ylitteiden lukumäärä noudattaa asymptoottisesti Poisson-jakaumaa, jolloin – iid satunnaismuuttujajonoa tarkastellessa – korkean tason ylitteiden *sattumisaikojen* voidaan odottaa noudattavan likimääräisesti Poisson-prosessia[15]. Kerrataan muutamat osiossa 1.4 esitetyt tulokset vielä tämän luvun kontekstissa. Olkoon $X_1, X_2, \ldots$ jono iid satunnaismuuttujia (tai satunnaismuuttujia stationaarisesta

[14] Ks. alaviite 7.
[15] Ks. liite C.

prosessista ääriarvoindeksillä $\theta = 1$), ja oletetaan että näiden kertymäfunktiolle pätee $F \in \mathrm{MDA}(H_\xi)$. Yhtälöiden (1.13) ja (1.14) perusteella tällöin pätee

$$\begin{aligned} \lim_{n\to\infty} n\ln(1-\bar{F}(c_n x + d_n)) &= \ln H_\xi(x), \\ \lim_{n\to\infty} n\bar{F}(c_n x + d_n) &= -\ln H_\xi(x). \end{aligned}$$

Kiinnostuksen kohteena ovat nyt kynnysjonon $(u_n(x))$ ylitteet, missä $u_n(x) = c_n x + d_n$ jollakin kiinteällä x:n arvolla. Kynnystason $u_n(x)$ ylityksien lukumäärä otoksessa $X_1, \ldots, X_n$ on binomijakautunut satunnaismuuttuja, $N_{u_n}(x) \sim \mathrm{Bin}(n, \bar{F}(u_n(x)))$, odotusarvona $\mathbb{E}(N_{u_n}(x)) = n\bar{F}(u_n(x))$. Ylitysten lukumäärä $N_{u_n}(x)$ suppenee nyt Poisson-jakautuneeseen satunnaismuuttujaan kun $n \to \infty$ (ks. s. 28). Koska (1.14) pätee, tapahtuu suppeneminen satunnaismuuttujaan $N_\infty \sim \mathrm{Poi}(\lambda(x))$, jonka odotusarvona on

$$\lambda(x) = -\ln(H_\xi(x)), \tag{1.26}$$

riippuen valitusta x.

Edellä mainitulla asetelmalla on läheinen yhteys myös GEV-jakaumaan. Olkoon N Poisson-jakautunut, $N \sim \mathrm{Poi}(\lambda)$, ja riippumaton iid satunnaismuuttujajonosta (X_n) GP-jakaumalla $G_{\xi,\beta}$. Merkitään $M_N = \max(X_1, \ldots, X_N)$. Tällöin

$$\mathbb{P}(M_N \le x) = \exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\} = H_{\xi,\mu,\sigma}(x), \tag{1.27}$$

missä $\mu = \frac{\beta}{\xi}(\lambda^\xi - 1)$ ja $\sigma = \beta\lambda^\xi$. Tulos siis sanoo, että kun ylitteiden lukumäärä on *täsmälleen* Poisson-jakautunut ja ylitteet noudattavat *täsmälleen* GP-jakaumaa, on ylitteiden maksimi täsmälleen GEV-jakautunut. Tulos (1.27) saadaan suoralla laskulla:

$$\begin{aligned} \mathbb{P}(M_N \le x) &= \sum_{n=0}^{\infty} e^{-\lambda}\frac{\lambda^n}{n!}G_{\xi,\beta}^n(x) = e^{-\lambda(1-G_{\xi,\beta}(x))} \\ &= \exp\left\{-\lambda\left(1+\xi\frac{x}{\beta}\right)^{-1/\xi}\right\} \\ &= \exp\left\{-\left(1+\xi\frac{x-\xi^{-1}\beta(\lambda^\xi-1)}{\beta\lambda^\xi}\right)^{-1/\xi}\right\}, \end{aligned}$$

kun $\xi \neq 0$. Tapauksessa $\xi = 0$ saadaan vastaavasti

$$P(M_N \le x) = \exp\left\{-e^{-\frac{x-\beta\ln\lambda}{\beta}}\right\}.$$

Kuten edellä mainittu, iid satunnaismuuttujajonolle ylitteiden lukumäärä on asymptoottisesti Poisson-jakautunut. Lauseen 1.47 mukaan ylitteiden suuruudet puolestaan noudattavat asymptoottisesti yleistettyä Pareto-jakaumaa (ylitejakauma on GP, kun $F \in \mathrm{MDA}(H_\xi)$). Tämä yhdistettynä tulokseen (1.27) motivoi mallintamaan korkean tason u ylitteitä pisteprosessina, jossa pisteet (t, x) kaksiulotteisessa avaruudessa kuvaavat ylityksien sattumista (t-dimensio tai -akseli) ja ylitteiden suuruuksia (x-dimensio). Todellakin, osoittautuu (ks. alaosio 1.6.1.5), että ylityksiin liittyvät pisteprossit suppenevat vastaavaan Poisson-pisteprosessiin.

Tarkastellaan seuraavaksi joitakin pisteprosesseihin liittyviä käsitteitä ja tuloksia. Tämän jälkeen esitetään iid dataa koskeva perusmalli, aikaulottuvuuden suhteen homogeeniseen Poisson-pisteprossiin perustuva ns. POT-malli, ja lopuksi yleistetään tämä ajan ja paikan suhteen epähomogeenisiin Poisson-pisteprosesseihin. Malliin liittyvää tilastollista päättelyä ja mallin estimointia käsitellään osiossa 2.6.

1.6.1 Pisteprosesseista

Pisteprosessi N on pisteiden X_i satunnainen jakauma avaruudessa. Tietylle pisteiden kokoelmalle $\{X_i : i \geq 1\}$ ja joukolle A laskee $N(A)$ niiden pisteiden lukumäärän, jotka osuvat joukkoon A. Etenkin sovellusten kannalta tärkeimpiä pisteprosesseja ovat prosessit, joille $N(A)$ on Poisson-jakautunut. Tämä johtaa Poisson-pisteprosessin tai Poisson-satunnaismitan määritelmään klassisen, välillä $[0, \infty)$ määritellyn, Poisson-prosessin yleistyksenä. Aloitetaan kiinnittämällä tarvittava notaatio. Seuraava esitys perustuu vahvasti lähteeseen Embrechts et. al. [2, luku 5].

Pisteprosessi voidaan hahmottaa intuitiivisesti seuraavalla tavalla. Olkoon E tila-avaruus jossa pisteet "elävät", ja tarkastellaan satunnaisvektoreiden jonoa (X_n) tila-avaruudessa E. Kun $A \subset E$, määritellään

$$N(A) = \#\{i : X_i \in A\} := \text{card}\{i : X_i \in A\},$$

missä $\text{card}(A)$ on joukon A mahtavuus eli kardinaliteetti (joukon "alkioiden lukumäärä"). $N(A)$ on siis joukkoon A osuvien pisteiden X_i lukumäärä, ja pisteprosessi on todennäköisyyskentällä $(\Omega, \mathscr{F}, \mathbb{P})$ määriteltyjen satunnaismuuttujien $N(\omega, A)$, $\omega \in \Omega$, kokoelma. Annetulle A siis $N(A) = N(\cdot, A)$ on satunnainen, ja (yleisten ehtojen vallitessa) $N(\omega, \cdot)$ määrittelee laskurimitan sopivalla tila-avaruuden E osajoukkojen muodostamalla sigma-algebralla $\mathscr{E}$. Vastaavasti kun valitaan tietty $\omega \in \Omega$, on $N(\omega, A)$ pisteiden lukumäärä joukossa A realisaatiossa ω.

Tässä esityksessä tila-avaruus E on aina äärellisulotteinen euklidinen avaruus $\mathbb{R}^d$, $d \geq 1$, tai sen osajoukko, ja intuition kannalta tila-avaruutta on helpointa ajatella tasona $\mathbb{R}^2$, vaikka yleinen teoria pätee paljon yleisemmissä avaruuksissakin. Tila-avaruus E on varustettu sigma-algebralla $\mathscr{E}$, joka otetaan vastaavasti Borel-σ-algebraksi, eli E:n avoimien joukkojen generoimaksi σ-algebraksi.

1.6.1.1 Määritelmä

Kun $x \in E$, määritellään mitta δ_x ehdosta

$$\delta_x(A) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A, \end{cases} \qquad \text{kun } \ A \in \mathscr{E}.$$

Tätä kutsutaan Dirac-mitaksi tai pistemassaksi. Olkoon $\{x_i : i \geq 1\}$ numeroituva joukko pisteitä tila-avaruudessa E. *Laskurimitta* (counting measure) E:ssä määritellään mittana m, joka on seuraavaa muotoa:

$$m(A) := \sum_{i=1}^{\infty} \delta_{x_i}(A) = \text{card}\{i : x_i \in A\}, \quad A \in \mathscr{E}.$$

Jos $m(K) < \infty$ kaikille kompakteille joukoille $K \subset E$ (eli m on *Radon*), kutsutaan mittaa m *pistemitaksi* (point measure). Olkoon $M_p(E)$ kaikkien joukossa E määriteltyjen pistemittojen avaruus varustettuna $M_p(E)$:n osajoukkojen generoimalla sigma-algebralla $\mathscr{M}_p(E)$.[16]

Pisteprosessi N avaruudessa E on mitallinen kuvaus todennäköisyysavaruudesta $(\Omega, \mathscr{F}, \mathbb{P}) \to (M_p(E), \mathscr{M}_p(E))$. Pisteprosessi on siis avaruuden $M_p(E)$ satunnainen elementti. Tyypillinen tapa määrittää pisteprosessi on seuraava: Olkoon $\{X_n : n \geq 1\}$ avaruuden E satunnaiselementtejä (tässä esityksessä käytännössä satunnaisvektoreita), ja asetetaan

$$N = \sum_{i=1}^{\infty} \delta_{X_i}. \tag{1.28}$$

Tällöin N on pisteprosessi, ja, jokaisella $\omega \in \Omega$,

$$N(\omega, A) = \sum_{i=1}^{\infty} \delta_{X_i(\omega)}(A), \tag{1.29}$$

$A \in \mathscr{E}$, määrittelee pistemitan $\mathscr{E}$:ssä.

Huomautus 1.48 (Yksinkertainen pisteprosessi) Olkoon $m = \sum_{i=1}^{\infty} \delta_{x_i}$ pistemitta avaruudessa E, ja olkoon $S_m = \{x \in E : m(\{x\}) \neq 0\}$. S_m muodostuu siis kokoelman $\{x_i, i \geq 1\}$ erillisistä pisteistä, ja se on pienin suljettu joukko A s.e. $m(A^c) = 0$ (siis m:n tukijoukko, engl. support). Jos $x \in S_m$, lukua $m(\{x\})$ kutsutaan pisteen x multiplisiteetiksi eli kerrannaisuudeksi, ja mittaa m kutsutaan *yksinkertaiseksi*, jos $m(\{x\}) \leq 1$ kaikilla $x \in E$. Vastaavasti pisteprosessi N on yksinkertainen, jos $\mathbb{P}(N(\{x\}) \leq 1, \forall\, x \in E) = 1$. Yksinkertaisen pisteprosessin realisaatiot ovat siis yksinkertaisia pistemittoja. Jos pistemitta tai pisteprosessi ei ole yksinkertainen, se on moninkertainen.

Vaihtoehtoisesti olkoon (y_i) jonon (x_i) osajono, joka sisältää kaikki (x_i):n erilliset arvot. y_i:n kerrannaisuus voidaan nyt kirjoittaa $n_i = \text{card}\{j : j \geq 1, y_i = x_j\}$, jolloin m saadaan muotoon

$$m = \sum_{i=1}^{\infty} n_i \delta_{y_i}.$$

Jos $n_i = 1$ kaikilla i, niin m on yksinkertainen.

Tarkastellaan havainnollistavia esimerkkejä. Seuraavat on otettu lähteestä [2, s. 223–225].

Esimerkki 1.49 (Uusiutumisprosessi)
Olkoon $(\xi_i)_{i\geq 1}$ *jono iid satunnaismuuttujia,* $\xi_i > 0$*, ja asetetaan* $T_n = \xi_1 + \cdots + \xi_n$*,* $n \geq 1$*. Jonon* (ξ_i) *generoimaa laskuriprosessia*

$$N(t) = \text{card}\{i : T_i \leq t\}, \quad t \geq 0$$

kutsutaan uusiutumisprosessiksi. Tähän prosessiin liittyy pisteprosessi

$$\tilde{N}(A) = \sum_{i=1}^{\infty} \delta_{T_i}(A), \quad A \in \mathscr{E},$$

[16] Tarkemmin, $\mathscr{M}_p(E)$ on pienin σ-algebra, joka sisältää kaikki muotoa $\{m \in M_p(E) : m(A) \in B\}$ olevat joukot mille tahansa $A \in \mathscr{E}$ ja Borel-joukolle $B \subset [0, \infty]$.

tila-avaruudella $E = [0, \infty)$. Kun otetaan joukoksi A väli $[0, t]$, $A = [0, t] \in \mathscr{E}$, saadaan

$$N(t) = \tilde{N}([0, t]), \quad t \geq 0.$$

Tämä nähdään vielä selvemmin kirjoittamalla pisteprosessi $\tilde{N}$ muotoon

$$N(A) = \sum_{i=1}^{\infty} \mathbb{1}_{\{T_i \in A\}} = \text{card}\{i : T_i \in A\},$$

missä $A \in \mathscr{E}$. Edellä esitetyssä mielessä jokaista uusiutumisprosessia vastaa jokin pisteprosessi. Näin määritelty pisteprosessi on yksinkertainen, sillä $0 < T_1 < T_2 < \cdots$ melkein varmasti. On yleisesti tunnettua (ks. liite C), että homogeenisen Poisson-prosessin odotusajat eli hyppyjen väliset ajat ovat iid eksponenttijakautuneita satunnaismuuttujia; ts. Poisson-prosessi on uusiutumisprossi eksponenttijakautuneilla ξ_i. Täten tavallinen Poisson-prosessi määrittelee "Poisson-pisteprosessin" (vrt. jäljempänä määriteltävä Poisson-satunnaismitta).

Esimerkki 1.50 (Uusiutumisprosessin indeksoima summa)
Tarkastellaan satunnaista summaa, jossa summattavien lukumäärä noudattaa edellisen esimerkin uusiutumisprosessia $(N(t))_{t \geq 0}$:

$$S(t) = \sum_{i=1}^{N(t)} X_i, \quad t \geq 0,$$

missä (X_i) on iid satunnaismuuttujien jono ja kaikilta osin riippumaton $(N(t))$:stä. Tällaiset mallit ovat tuttuja riskiteoriasta, jossa tyypillisesti X_i edustaa satunnaisella hetkellä T_i sattuvan vahingon suuruutta, ja $S(t)$ on hetkeen t mennessä sattuneiden vahinkojen summa eli kokonaisvahinkomäärä. Ks. esim. [16]. Eräs $(S(t))$:hen liittyvä yksinkertainen pisteprosessi on

$$\tilde{N}(A) = \sum_{i=1}^{\infty} \delta_{(T_i, X_i)}(A), \quad A \in \mathscr{E}$$

tila-avaruudessa $E = [0, \infty) \times \mathbb{R}$. Esimerkiksi kun otetaan $A = (a, b] \times (u, \infty) \in \mathscr{E}$,

$$\tilde{N}((a, b] \times (u, \infty)) = \text{card}\{i : a < T_i \leq b, X_i > u\}$$

laskee välillä $(a, b]$ sattuvien ja tason u ylittävien tapahtumien lukumäärän. Yllä määritelty $\tilde{N}$ muistuttaa läheisesti jäljempänä käsiteltävää ylitysten pisteprosessia.

Määritelmä 1.51 (Merkitty pisteprosessi) *Olkoon $N = \sum_i \delta_{X_i}$ pisteprosessi avaruudella E, ja olkoon E' toinen avaruus.* Merkitty pisteprosessi *alla olevalla prosessilla N on mikä tahansa pisteprosessi*

$$\tilde{N} = \sum_{i=1}^{\infty} \delta_{(X_i, Z_i)}$$

avaruudella $E \times E'$. E':n satunnaiselementtiä Z_i kutsutaan X_i:hin liittyväksi merkiksi.

Tietyssä mielessä merkitty pisteprosessi $\tilde{N}$ on vain pisteprosessi tuloavaruudessa $E \times E'$. Tilanteesta riippuen voi olla edullisempaa tarkastella tiettyä pisteprosessia joko merkittynä pisteprosessina tai pisteprosessina tuloavaruudessa.

1.6.1.2 Pisteprosessin jakauma

Pisteprosessin N realisaatiot ovat pistemittoja $N(\omega, \cdot)$. Täten pisteprosessin jakauma on määritelty pistemittojen muodostamassa avaruudessa $\mathscr{M}_p(E)$:

Määritelmä 1.52 (Pisteprosessin jakauma)
Olkoon annettu pisteprosessi N määriteltynä todennäköisyyskentällä $(\Omega, \mathscr{F}, \mathbb{P})$. N:n jakauma *P_N on todennäköisyysmitta $(M_p(E), \mathscr{M}_p(E))$:llä, missä $P_N = \mathbb{P} \circ N^{-1} = \mathbb{P}(N \in \cdot\,)$.*

Jakaumaa $P_N(A) = \mathbb{P}(N \in A)$, $A \in \mathscr{M}_p(E)$ itsessään ei ole helppo mieltää, mutta osoittautuu, että pisteprosessin jakauman määräävät yksikäsitteisesti sen äärellisulotteiset jakaumat. Äärellisulotteisilla jakaumilla tarkoitetaan satunnaisvektoreiden $(N(A_1), \ldots, N(A_m))$ jakaumien

$$\mathbb{P}\left(N(A_1) = n_1, \ldots, N(A_m) = n_m\right), \quad n_i \geq 0, \quad i = 1, \ldots, m,$$

kaikilla $A_1, \ldots, A_m \in \mathscr{E}$, $m \geq 1$, muodostamaa perhettä (ks. liite D). Jakauman P_N määräämiä äärellisulotteisia jakaumia on huomattavasti helpompi käsitellä kuin jakaumaa P_N itse.

Määritellään seuraavaksi keskeisen pisteprosessien luokan muodostavat Poisson-pisteprosessit.

1.6.1.3 Poisson-pisteprosessi

Olkoon μ Radon-mitta $\mathscr{E}$:llä (eli $\mu(K) < \infty$ kompakteille joukoille $K \subset E$).

Määritelmä 1.53 (Poisson-satunnaismitta (PRM))
Pisteprosessia N kutsutaan Poisson-pisteprosessiksi *tai* Poisson-satunnaismitaksi keskiarvomitalla μ *(merkitään lyhyesti* $\mathrm{PRM}(\mu)$*), jos seuraavat kaksi ehtoa täyttyvät:*

(a) Mille tahansa $A \in \mathscr{E}$ ja ei-negatiiviselle kokonaisluvulle $k \geq 0$,

$$\mathbb{P}(N(A) = k) = \begin{cases} e^{-\mu(A)} \dfrac{(\mu(A))^k}{k!}, & \text{jos} \quad \mu(A) < \infty, \\ 0, & \text{jos} \quad \mu(A) = \infty. \end{cases}$$

(b) Mille tahansa $k \geq 1$, jos $A_1, \ldots, A_k$ ovat keskenään erillisiä joukkoja ($A_i \cap A_j = \emptyset, i \neq j$) $\mathscr{E}$:ssä, niin

$$N(A_1), \ldots, N(A_k)$$

ovat riippumattomia satunnaismuuttujia.

Määritelmän 1.53 mukainen Poisson-satunnaismitta on olemassa (ks. liite D.1), ja sen jakauma määräytyy yksikäsitteisesti ehdoista (a) ja (b) yllä. N:llä on siis riippumattomat lisäykset, ja jokaisella A pisteiden satunnainen määrä joukossa A noudattaa Poisson-jakaumaa odotusarvolla $\mathbb{E}(N(A)) = \mu(A)$.

Esimerkki 1.54 (Homogeeninen Poisson-satunnaismitta)
Historiallisesti ensimmäisinä tutkitut ja hallitut Poisson-prosessit ovat tavallisia

Poisson-prosesseja positiivisella reaaliakselilla, $E = [0, \infty)$ (ks. liite C). Näille keskiarvomitta on

$$\mu(A) = \lambda |A| = \lambda \int_A dx, \quad A \in \mathscr{E},$$

missä $|\cdot|$ on Lebesgue-mitta $[0, \infty)$:llä, ja λ on intensiteetti.

Olkoon nyt N $\mathrm{PRM}(\lambda |\cdot|)$ tila-avaruudessa $E \subset \bar{\mathbb{R}}^d$, $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, missä $\lambda > 0$ ja $|\cdot|$ on E:n Lebesgue-mitta. Homogeenisen Poisson-prosessin $[0, \infty)$:llä yleistyksenä N on homogeeninen PRM tai homogeeninen Poisson-pisteprosessi intensiteetillä λ. Jos lisäksi PRM:n keskiarvomitta on absoluuttisesti jatkuva Lebesgue-mitan suhteen (eli $|\cdot| = 0 \Rightarrow \mu = 0$), jolloin on olemassa ei-negatiivinen funktio $\lambda(\cdot)$ s.e.

$$\mu(A) = \int_A \lambda(x) dx, \quad A \in \mathscr{E},$$

niin $\lambda(\cdot)$ on PRM:n intensiteetti(funktio).

1.6.1.4 Pisteprosessien suppeneminen jakauman suhteen

Pisteprosessien suppeneminen jakauman suhteen tai heikko suppeneminen on perustyökalu mm. ääriarvoteoriaan liittyviä asymptoottisia tuloksia käsiteltäessä. Ei ole kuitenkaan välittömästi ilmeistä, mitä pisteprosessien heikko suppeneminen itse asiassa tarkoittaa, eikä täsmällisiä suppenemistuloksia voida muotoilla ilman tämän esityksen laajuuden ja rajauksen ylittäviä käsitteitä ja tuloksia heikon suppenemisen teoriasta metrisissä avaruuksissa. Seuraavan yleiskatsauksen tarkoituksena on antaa intuitiivinen käsitys pisteprosessien heikosta suppenemisesta, ja esittää myöhemmin tarvittavat tulokset. Esitys seuraa lähdettä [2, kappale 5.2 ja liite A.2], rajoittuen kuitenkin intuitiivisiin perusteluihin; hieman tarkempi pisteprosessien heikon suppenemisen käsittely on delegoitu liitteeseen D.

Huomautetaan ensin, että suppenemisesta puhumiseksi täytyy määritellä metriikka jonka suhteen suppeneminen määritellään. Sopivaksi suppenemiskäsitteeksi avaruudessa $M_p(E)$ osoittautuu ns. *epämääräinen* suppeneminen (vague convergence, merk. $\xrightarrow{v}$). Epämääräinen topologia on metrisoitavissa niin, että $M_p(E)$:stä (*epämääräisellä* metriikalla varustettuna) tulee täydellinen, separoituva metrinen avaruus (ks. [17, luku 1.3] tai [3, luku 3.5]). Tällöin voidaan puhua todennäköisyysmittojen heikosta suppenemisesta $(M_p(E), \mathscr{M}_p(E))$:llä, missä $\mathscr{M}_p(E)$ on epämääräisen topologian synnyttämä Borel-σ-algebra.

Otetaan nyt tarkastelun kohteeksi pisteprosessit $N, N_1, N_2, \ldots$ tila-avaruudessa $E \subset \bar{\mathbb{R}}^d$. Edeltä muistetaan, että näiden pisteprosessien jakauma $M_p(E)$:ssä määrittyy niiden äärellisulotteisten jakaumien kautta. Täten luonnolliselta tuntuva vaatimus jonon (N_n) heikolle suppenemiselle N:ään on, että mille tahansa "hyville" $\mathscr{E}$:n Borel-joukoille $A_1, \ldots, A_m$ ja mille tahansa kokonaisluvulle $m \geq 1$ pätee

$$\mathbb{P}(N_n(A_1), \ldots, N_n(A_m)) \to \mathbb{P}(N(A_1), \ldots, N(A_m))), \tag{1.30}$$

kun $n \to \infty$. Toisaalta jokainen pisteprosessi N voidaan nähdä stokastisena prosessina, eli tässä joukon $A \in \mathscr{E}$ indeksoimana satunnaismuuttujien $N(A)$ kokoelmana. N on siis avaruuden $M_p(E)$ ääretönulotteinen elementti. Äärellisulotteis-

ten jakaumien suppeneminen ei yleisesti ole riittävä stokastisten prosessien suppenemiseksi. Suppenemisen varmistamiseksi tarvitaan lisäehtoja, ja sopivaksi pisteprosessien yhteydessä osoittautuu *tiukkuus*[17] (tightness). Ehto varmistaa, että stokastisten prosessien todennäköisyysmassa ei karkaa tila-avaruuden "hyvistä" (kompakteista) joukoista suppenemisprosessin yhteydessä, kun $n \to \infty$. Ks. [3, luku 3.5].

Pisteprosesseilla osoittautuu olevan hyvin hyödyllinen ominaisuus: pisteprosessien jono on tiukka jos ja vain jos kaikki sen yksiulotteiset jakaumat ovat tiukkoja. Täten tiukkuus seuraa pisteprosessien äärellisulotteisten jakaumien suppenemisesta, ja saadaan seuraava määritelmä:

Määritelmä 1.55 (Pisteprosessien heikko suppeneminen)
Olkoon (N_n), N pisteprosesseja tila-avaruudessa $E \subset \bar{\mathbb{R}}^d$ varustettuna Borel-joukkojen σ-algebralla $\mathscr{E}$. Sanotaan, että (N_n) suppenee heikosti *N:ään $M_p(E)$:ssä,*

$$N_n \xrightarrow{d} N, \quad n \to \infty,$$

jos

$$(N_n(A_1), \ldots, N_n(A_m)) \xrightarrow{d} (N(A_1), \ldots, N(A_m)), \quad n \to \infty,$$

kaikilla $m \geq 1$ ja kaikilla $A_1, \ldots, A_m \in \mathscr{E}$ s.e. $\mathbb{P}(N(\partial A_i) = 0) = 1$, $i = 1, \ldots, m$. (∂A tarkoittaa joukon A reunaa.)

Huomautus 1.56 (Poisson-pisteprosessin suppeneminen)
Olkoon $N, N_1, N_2, \ldots$ Poisson-pisteprosesseja keskiarvomitoilla $\mu, \mu_1, \mu_2, \ldots$. Voidaan osoittaa, että $N_n \xrightarrow{d} N$ jos ja vain jos $\mu_n \xrightarrow{v} \mu$ (epämääräisesti). Ks. myös liite D.

Huomautus 1.57 (Uusiutumisprosessien suppeneminen)
Jos (N_n) on jono uusiutumisprosesseja ja $N_n \xrightarrow{d} N$, niin N on myös uusiutumisprosessi ([17, s. 16]).

1.6.1.5 Ylitteiden pisteprosessi

Käsitellään seuraavaksi ääriarvoteoriaan läheisesti liittyvää ylitteiden pisteprosessia ([2, luku 5.3]). Olkoon u_n reaaliluku ja (X_n) jono satunnaismuuttujia. *Ylitteiden pisteprosessi*

$$N_n(\cdot) = \sum_{i=1}^{n} \delta_{\frac{i}{n}}(\cdot) \mathbb{1}_{\{X_i > u_n\}}, \quad n = 1, 2, \ldots, \tag{1.31}$$

tila-avaruudella $E = (0, 1]$ laskee tällöin kynnyksen u_n ylityksien lukumäärän jonossa $X_1, \ldots, X_n$. Esimerkiksi kun otetaan koko väli $(0, 1]$, on

$$N_n((0,1]) = \text{card}\{i : 0 < n^{-1} i \leq 1 \ \wedge \ X_i > u_n\} = \text{card}\{i \leq n : X_i > u_n\}.$$

Ylitteiden pisteprosessi voidaan vaihtoehtoisesti kirjoittaa muodossa

$$N_n(\cdot) = \sum_{i=1}^{n} \delta_{(\frac{i}{n}, X_i)}(\cdot), \quad n = 1, 2, \ldots, \tag{1.32}$$

[17] Todennäköisyysmittojen P_n jono metrisessä (täydellisessä, separoituvassa) avaruudessa on tiukka, jos kaikilla $\varepsilon > 0$ on olemassa kompakti joukko K_ε s.e. $P_n(K_\varepsilon) > 1 - \varepsilon$ kaikilla n.

kaksiuloitteisella tila-avaruudella $E = (0,1] \times (u_n, \infty)$.

Huomautus 1.58 (Empiirinen prosessi) *Kun $X_1, \ldots, X_n$ ovat avaruuden E iid satunnaiselementtejä, muotoa*

$$N = \sum_{i=1}^{n} \delta_{X_i}$$

olevaa pisteprosessia kutsutaan yleisesti (n-otoksen) empiiriseksi prosessiksi *E:llä.*

Pisteprosessin läheiset yhteydet ääriarvoteoriaan nähdään helposti. Merkitään otoksen $(X_i)_{i=1}^n$ k:nneksi suurinta järjestystunnuslukua $X_{k,n}$. Tällöin

$$\begin{aligned}\{N_n((0,1]) = 0\} &= \{\text{card}\{i \leq n : X_i > u_n\} = 0\} \\ &= \{\max(X_1, \ldots, X_n) \leq u_n\} \\ &= \{M_n \leq u_n\},\end{aligned}$$

$$\begin{aligned}\{N_n((0,1]) < k\} &= \{\text{card}\{i \leq n : X_i > u_n\} < k\} \\ &= \{\text{vähemmän kuin } k \text{ muuttujista } X_i, i \leq n, \text{ ylittää tason } u_n\} \\ &= \{X_{k,n} \leq u_n\}.\end{aligned}$$

N_n:n katsotaan tässä olevan elementti pisteprosessien jonossa (N_n), ja kiinnostuksen kohteena on pisteprosessin asymptoottinen käyttäytyminen, eli käyttäytyminen kun $n \to \infty$. Tämän osion alusta (ks. 1.6) muistetaan, että tason u_n ylitysten lukumäärä kiinteässä otoksessa noudattaa asymptoottisesti Poisson-jakaumaa. Ylitysten pisteprossille (1.31) pätee analoginen tulos: voidaan osoittaa, että pisteprosessien jono (N_n) suppenee heikosti (homogeeniseen) Poisson-pisteprosessiin N.

Olkoon nyt (X_n) iid satunnaismuuttujia (yhteisellä jakaumalla F) ja (u_n) jono kynnystasoja, $u_n \in \mathbb{R}$. Propositiosta (1.40) muistetaan, että relaatio $\mathbb{P}(M_n \leq u_n) \to \exp(-\tau)$, $\tau \in [0, \infty]$, pätee jos ja vain jos

$$n\bar{F}(u_n) = \mathbb{E}\sum_{i=1}^{n} \mathbb{1}_{\{X_i > u_n\}} \to \tau. \tag{1.33}$$

Propositio 1.59 (Ylitteiden pisteprosessin heikko suppeneminen) *Olkoon (X_n) jono iid satunnaismuuttujia jakaumalla F, ja olkoon (u_n) jono kynnysarvoja, s.e.* (1.33) *pätee jollakin $\tau \in (0, \infty)$. Tällöin ylitteiden pisteprosessien* (1.31) *jono (N_n) suppenee $M_p(E)$:ssä heikosti pisteprosessiin N,*

$$N_n \xrightarrow{d} N, \quad n \to \infty,$$

missä N on homogeeninen Poisson-prosessi tila-avaruudella E intensiteetillä τ; ts. N on $\text{PRM}(\tau\,|\cdot|)$*, missä $|\cdot|$ tarkoittaa Lebesgue-mittaa E:llä.*

Todistusta varten ks. liite D. Sovelluksissa tarkastellaan jatkossa tyypillisesti väliä $E = (0, t]$, $0 < t < \infty$, mikä tulkitaan aikaulottuvuudeksi. Tila-avaruudelle $E = (0, 1]$ muotoillut tulokset laajenevat välittömästi tälle.

Olkoon kynnystasojen jono $u_n = u_n(x)$ muotoa $u_n(x) = c_n x + d_n$ jollakin kiinteällä $x \in \mathbb{R}$, kuten aiemmin. Raja-arvona olevan Poisson-pisteprosessin keskiarvomitaksi saadaan alaosion 1.6.1.3 ja (1.26):n perusteella

$$\mu(A) = \int_{t_1}^{t_2} \lambda(x)\,\mathrm{d}t = (t_2 - t_1)\lambda(x) = -(t_2 - t_1)\ln H_\xi, \tag{1.34}$$

kun $A = (t_1, t_2) \subset \mathscr{E}$, missä siis $\lambda(x) = -\ln H_\xi(x)$. Tässä siis rajaprosessin intensiteetti ei riipu ajasta ja ottaa (valitulla x) vakioarvon $\lambda(x) =: \lambda$.

1.6.2 POT-malli

Edellä ja osiossa 1.5 esitetty ohjaa kuvaamaan ylitteitä säännöllisin välein havaitussa (regularly spaced) iid datassa[18] seuraaviin oletuksiin perustuvalla asymptoottisella mallilla:

- Ylitykset tapahtuvat ajassa homogeenisen Poisson-prosessin mukaisesti.
- Ylitteiden suuruudet ovat iid satunnaismuuttujia ja riippumattomia ylitysajoista.
- Ylitteiden suuruudet noudattavat yleistettyä Pareto-jakaumaa.

Tätä kutsutaan Peaks-Over-Threshold –malliksi (POT); ks. [23], [5, luku 7.4]. Mallia voidaan lähestyä useasta eri näkökulmasta: se voidaan tulkita myös merkityksi Poisson-pisteprosessiksi, missä ylitysajat ovat pisteitä ja GP-jakautuneet ylitteet merkkejä. Erityisen hyödylliseksi osoittautuu mallin esitys kaksiulotteisena Poisson-pisteprosessina, jossa pisteet (t, x) kaksiulotteisessa tila-avaruudessa kuvaavat ylitysten sattumisaikaa ja ylitteiden suuruutta. Tarkastellaan tätä esitystä seuraavassa.

Oletetaan, että data koostuu iid satunnaismuuttujajonosta $(X_t)_{t=1}^n$, ja kiinnitetään korkea kynnystaso u, jolloin tason ylityksien lukumäärä on N_u. Ylitteet muodostavat nyt datan $\{(t, X_t) : 1 \le t \le n, X_t > u\}$; numeroidaan nämä vaihtoehtoisesti peräkkäin käyttäen notaatiota $\{(T_j, \tilde{X}_j) : j = 1, \ldots, N_u\}$. Ylityksien ajat ilmaistaan siis nyt aikasarjan luonnollisella aikaskaalalla (ts. t_i välillä $(0, n]$ eikä $n^{-1}t_i$ välillä $(0, 1]$).

Oletetaan, että pisteet tila-avaruudessa $E = (0, n] \times (u, \infty)$ sattuvat Poisson-pisteprosessin $N(\cdot)$ mukaisesti, missä

$$N(A) = \sum_{t=1}^{n} \delta_{(t,X_t)}(A) = \sum_{t=1}^{n} \mathbb{1}_{\{(t,X_t)\in A\}}, \quad A \in \mathscr{E},$$

intensiteetillä

$$\lambda(t, x) = \frac{1}{\sigma}\left(1 + \xi\frac{x - \mu}{\sigma}\right)^{-1/\xi - 1},$$

kun $1 + \xi(x - \mu)/\sigma > 0$, ja $\lambda(t, x) = 0$ muulloin. Tämä tarkoittaa mm., että ylitteiden lukumäärät erillisissä suorakulmioissa $A_1, A_2, \ldots$ ovat riippumattomia Poisson-jakautuneita satunnaismuuttujia keskiarvoinaan $\mu(A_1), \mu(A_2), \ldots$.

[18] Sama pätee jälleen myös stationaarisen prosessin generoimaan dataan, kun prosessin ääriarvoindeksi $\theta = 1$.

Käytetään keskiarvomitasta jatkossa nimitystä intensiteettimitta, ja merkitään tätä $\mu(\cdot) = \Lambda(\cdot)$.

Intensiteetti yllä ei riipu ajasta t mutta riippuu "paikasta" x, ja kaksiulotteinen Poisson-prosessi $N(\cdot)$ on siis epähomogeeninen. Merkitään intensiteettiä lyhyemmin $\lambda(x) := \lambda(t, x)$, $\forall t$. Prosessin intensiteettimitta Λ on nyt

$$\Lambda(A) = \int_{t_1}^{t_2} \int_x^\infty \lambda(y)\, \mathrm{d}y\, \mathrm{d}t = -(t_2 - t_1) \ln H_{\xi,\mu,\sigma}(x) \\ = (t_2 - t_1) \left(1 + \xi \frac{x - \mu}{\sigma}\right)^{-1/\xi},$$

muotoa $A = (t_1, t_2) \times (x, \infty) \in \mathscr{E}$ olevilla joukoilla.

Edellisestä seuraa, että tason $x \geq u$ ylityksien yksiulotteinen prosessi (kiinnitetyllä x) on homogeeninen Poisson-prosessi intensiteetillä

$$\tau(x) := -\ln H_{\xi,\mu,\sigma}(x) = \left(1 + \xi \frac{x - \mu}{\sigma}\right)^{-1/\xi}.$$

Tarkastellaan seuraavaksi tason u ylitteiden suuruuksia. Edustakoon X geneeristä satunnaismuuttujaa jonosta (X_t). Tason u ylitteiden jakauma eli ylitejakauma F_u saadaan tasojen u ja $u + x$ ylitysintensiteettien suhteena:

$$\bar{F}_u(x) = \mathbb{P}(X > u + x | X > u) = \frac{\tau(u + x)}{\tau(u)} \\ = \left(1 + \frac{\xi x}{\sigma + \xi(u - \mu)}\right)^{-1/\xi} = \bar{G}_{\xi,\beta}(x),$$

missä $\beta = \sigma + \xi(u - \mu) > 0$. Mallin implikoima ylitteiden suuruuksien jakauma, kun ylitys tapahtuu, on siis GP-jakauma.

Nähdään myös, että kaksiulotteinen Poisson-pisteprosessimalli implikoi ylitteiden maksimin olevan (asymptoottisesti) GEV-jakautunut: Tarkastellaan tapahtumaa $\{M_n \leq x\}$ jollakin $x \geq u$, missä $M_n = \max(X_1, \ldots, X_n)$, kuten aiemmin. Nyt $\{M_n \leq x\} = \{N((0, n] \times (x, \infty)) = 0\}$ (ei pisteitä joukossa $A = (0, n] \times (x, \infty)$), joten $\mathbb{P}(M_n \leq x) = \mathbb{P}(N(A) = 0) = \exp(-\Lambda(A)) = H_{\xi,\mu,\sigma}(x)$, mikä on juuri blokkimaksimimenetelmän GEV-malli maksimeille; vrt. myös tulokseen (1.27).

1.6.3 Yleisemmät mallit

Edellisen osion POT-mallissa Poisson-prosessin intensiteetti ei ollut aikariippuvainen,

$$\lambda(t, x) = \frac{1}{\sigma} \left(1 + \xi \frac{x - \mu}{\sigma}\right)^{-1/\xi - 1}.$$

Suoraviivaisin tapa yleistää POT-malli on asettaa malliparametrit (ξ, μ, σ) riippumaan ajasta (tai muista selittävistä muuttujista), jolloin intensiteetiksi pisteessä $(t, x) \in E$ tulee

$$\lambda(t, x) = \frac{1}{\boldsymbol{\sigma}(t)} \left(1 + \boldsymbol{\xi}(t) \frac{x - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)}\right)^{-1/\boldsymbol{\xi}(t) - 1}, \tag{1.35}$$

kun $1 + \boldsymbol{\xi}(t)(x - \boldsymbol{\mu}(t))/\boldsymbol{\sigma}(t) > 0$, ja $\lambda(t, x) = 0$ muulloin. Tuloksena saatavan epähomogeenisen kaksiulotteisen Poisson-prosessin intensiteettimitta on siis, muotoa $A = (t_1, t_2) \times (x, \infty) \in \mathscr{E}$ olevilla joukoilla,

$$\begin{aligned}
\Lambda(A) &= \int_{t_1}^{t_2} \int_{x}^{\infty} \lambda(t, y) \, \mathrm{d}y \, \mathrm{d}t \\
&= \int_{t_1}^{t_2} \int_{x}^{\infty} \frac{1}{\boldsymbol{\sigma}(t)} \left(1 + \boldsymbol{\xi}(t) \frac{x - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)}\right)^{-1/\boldsymbol{\xi}(t) - 1} \mathrm{d}y \, \mathrm{d}t \\
&= -\int_{t_1}^{t_2} \ln H_{\boldsymbol{\xi}(t), \boldsymbol{\mu}(t), \boldsymbol{\sigma}(t)}(x) \, \mathrm{d}t \\
&= \int_{t_1}^{t_2} \left(1 + \boldsymbol{\xi}(t) \frac{x - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)}\right)^{-1/\boldsymbol{\xi}(t)} \mathrm{d}t.
\end{aligned}$$

Kuten edellisessä osiossa, nähdään että edellä määritelty pisteprosessimalli implikoi ylitteiden suuruuksien olevan GP-jakautuneita (jakaumalla $G_{\boldsymbol{\xi}(t), \boldsymbol{\beta}(t)}(x) = G_{\boldsymbol{\xi}(t), \boldsymbol{\sigma}(t) + \boldsymbol{\xi}(t)(u - \boldsymbol{\mu}(t))}(x)$ pisteessä $(t, x) \in E$).

1.7 Historiaa

Ääriarvoteoria on klassinen tutkimusalue todennäköisyysteorian ja matemaattisen tilastotieteen piirissä. Tutkimusalan voidaan muodollisesti katsoa alkaneen vuosien 1927–1928 tienoilla. Julkaisusta Fréchet ([18], 1927) löytyy kaksi kolmesta ääriarvojakaumasta (jotka nykyään tunnetaan nimillä Fréchet ja Weibull); kaikki kolme on esitetty klassisessa paperissa Fisher ja Tippett ([19], 1928). Ääriarvojakaumat on nimetty seuraavien henkilöiden mukaan: Maurice Fréchet (1878–1973), ranskalainen matemaatikko; Emil Gumbel (1891–1966), saksalainen tilastotieteilijä; ja Waloddi Weibull (1887–1979), ruotsalainen insinööri. Gnedenko ([20], 1943) todisti ensimmäisenä maksimien rajajakaumia koskevat tulokset matemaattisesti tarkasti.

Yleistetyn Pareto-jakauman määritelmä on alunperin peräisin Pickandsilta ([21], 1975). Heikko raja-arvoteoria ylitejakaumille on saanut alkunsa julkaisusta Balkema ja de Haan ([22], 1974).

Kaksiulotteista Poisson-pisteprosessimallia ääriarvojen kuvaamiseen käytännössä sovellettiin ensimmäisenä julkaisussa Smith ([23], 1989).

Tarkastellaan seuraavaksi tarkemmin erästä traagista tapahtumaa, jolla on ollut suuri merkitys ääriarvoteorian historiassa. Esimerkki toimii samalla myös motivaationa seuraavan luvun tarkasteluihin – ja on sellaisenaan mielenkiintoinen.

1.7.1 Pohjanmeren tulva

Luultavasti merkittävin yksittäinen ääriarvoteorian ja erityisesti sen tilastollisten sovelluksien kehittymiseen vaikuttanut ulkoinen tekijä on ollut tammikuun 31. päivän iltana vuonna 1953 sattunut Pohjanmeren tulviminen. Kuten

Bingham ([24]) muotoilee, tuolloin todellisuus otti kiinni ja ohitti matematiikan, "mathematics was overtaken by reality". Myrskyaallokko ja sen aiheuttama laaja tulviminen johti 1 800 ihmisen kuolemaan Hollannissa, ja Alankomaiden hallitus asetti välittömästi tapahtuman seurauksena korkean prioriteetin sattuneen kaltaisten tulvaonnettomuuksien syiden ymmärtämiselle ja onnettomuuksien ehkäisemiselle. Merenpinnan alapuolella sijaitsevan Alankomaiden kohdalla ongelmassa oli siis kyse siitä, kuinka määrätä turvalliset padonkorkeudet, joita vesi ei ylitä riittävän suurella varmuudella, ottaen kuitenkin huomioon taloudelliset realiteetit ja kustannus-hyöty–suhteen.

Koska patojen kohdalla uhkana ovat nimenomaan aallonkorkeuden maksimiarvot, ääriarvoteoria on välittömästi relevantti ongelman ensimmäisen osan, eli vedenkorkeuden mallintamisen osalta. Hieman tarkemmin muotoiltuna, matemaattinen ongelma on seuraava: annettuna pieni luku p (luokkaa 10^{-4} tai 10^{-3}), on määrättävä padon korkeus siten, että tulvan todennäköisyys (padon ylittävän vedenkorkeuden sattumistodennäköisyys) tiettynä vuonna on p. Kyseessä on siis olennaisesti vedenkorkeuden todennäköisyysjakauman korkeiden kvantiilien estimointiongelma. Stationaarisessa tapauksessa (kun jakauma ei muutu ajan suhteen) edellä mainittua p:n arvoa vastaava $(1 - 10^{-4})$-kvantiili vastaa tapahtumaa, joka sattuu kerran 10 000 vuodessa. Tätä kutsutaan 10 000-vuoden toistumistasoksi (ks. tarkemmin alaosio 2.2.2 jäljempänä).

Kuten mainittu, hollantilaisten matemaatikkojen työ van Dantzigin alaisuudessa Alankomaiden patoprojektissa edisti ääriarvoteorian tilastollisten menetelmien kehitystä ja sen käytännön sovelluksia suuresti. Suoritetuilla tilastollisilla analyyseillä oli myös huomattava merkitys lopullisessa patojen korkeutta koskeneessa päätöksenteossa. Alankomaiden ns. suistolakiin (Deltawet, 8.5.1958) kirjoitettiin vaatimus, että hallituksen täytyy pitää tulvakatastrofin riski tietyissä rajoissa, ja tarvittaessa ryhtyä toimiin riskin alentamiseksi mikäli ilmiöstä saadaan uutta tietoa; hyväksyttäväksi riskiksi meriveden aiheuttamalle tulvimiselle määriteltiin alueesta riippuen tulvan sattuminen kerran 10 000, 4 000 tai 2 000 vuodessa. Rajat on sisällytetty myös uuteen, 22.12.2009 voimaan tulleeseen vesilakiin (Waterwet).[19]

Merenpinnan korkeus ilmoitetaan Alankomaissa tyypillisesti suhteessa keskivedenkorkeuteen N.A.P. (Normaal Amsterdams Peil). Vuoden 1953 tulvan aiheutti (N.A.P. + 3.85) metrin korkuinen myrskyaallokko, kun suurimman kirjoitetuissa lähteissä mainitun, vuoden 1570 tulvan aiheuttaneen aallokon korkeudeksi on arvioitu (N.A.P. + 4) m. Van Dantzig –raportti estimoi vedenkorkeuden vuosittaisen maksimin $(1 - 10^{-4})$-kvantiiliksi eli kerran kymmenessä tuhannessa vuodessa sattuvaksi tapahtumaksi (N.A.P. + 5.14) m.

Yhtenä tuloksena ääriarvoteorian noususta enemmän käytännön merkitystä – ja siten suuremman prioriteetin – omaavaksi matematiikan alaksi Hollannissa voidaan nähdä myös muuan Laurens de Haanin väitöskirja [25], ja sitä seurannut merkittävä tutkimus. Kuvailevasti nimetyssä paperissaan [26], "Fighting the Arch-Enemy with Mathematics"[20], de Haan kuvaa Alankomaiden patoprojektin yhteydessä tehtyä teoreettista ja soveltavaa työtä tarkemmin.

[19] Ks. `http://en.wikipedia.org/wiki/Delta_Works` ja `http://nl.wikipedia.org/wiki/Deltawet_(2011)`.

[20] Siteerattu teoksessa [2]; kirjoittaja ei yrityksistä huolimatta onnistunut saamaan de Haanin paperia [26] käsiinsä.

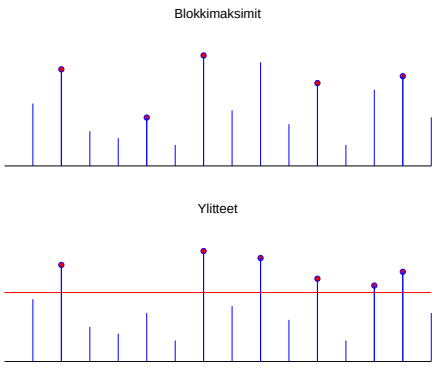
Lisää historiaa sekä viittaukset löytyvät esimerkiksi teoksesta [2], josta edellinen esimerkki on mukailtu.

Luku 2

Tilastolliset menetelmät

Tarkastellaan seuraavaksi edellisessä luvussa esitettyyn teoriaan perustuvia tilastollisia menetelmiä. Luvun esitys perustuu lähteisiin [1] ja [5].

Yleisimmät lähestymistavat äärimmäisten ilmiöiden tilastolliseksi kuvaamiseksi ovat ns. blokkimaksimimenetelmä ja ylitemenetelmä. Kuvassa 2.1 on havainnollistettu menetelmien eroa datan käytön kannalta: Ensimmäisessä havaintojoukko jaetaan n:n havainnon blokkeihin, ja blokkimaksimit mallinnetaan yleistetyllä ääriarvojakaumalla (GEV), kun taas jälkimmäisessä kaikkiin tietyn korkean tason ylittäviin havaintoihin (ylitteisiin) sovitetaan yleistetty Pareto-jakauma (GPD). Jälkimmäisen lähestymistavan yleistys on tarkastella pisteprosesseja, jotka kuvaavat sekä ylitteiden suuruudet että niiden sattumisen ajassa.



Kuva 2.1: Blokkimaksimimenetelmä vs. ylitemenetelmä.

Luku on organisoitu seuraavasti. Ensin tarkastellaan blokkimaksimimenetelmää ja GEV-jakauman sovittamista dataan suurimman uskottavuuden (SU) menetelmää käyttäen (osio 2.2). Osiossa 2.3 käsitellään ylitemenetelmää ja GP-jakauman sovittamista SU-menetelmällä. Yksinkertaisuuden vuoksi dataa pidetään realisaationa iid prosessista, tai stationaarisesta prosessista ääriarvoindeksillä $\theta = 1$: stationaariseen prosessiin ääriarvoindeksillä $\theta < 1$ liittyvän havainto-

jen kasaantumisen huomiointia käsitellään lyhyesti osiossa 2.4.[1] Epästationaarisia prosesseja käsitellään osiossa 2.5. Lopuksi osiossa 2.6 tarkastelu yleistetään pisteprosesseihin, joita käyttäen menetelmien yleistys epästationaariseen tapaukseen onnistuu luonnollisella tavalla.

Aivan keskeinen osa tilastollista mallinnusta on tarkastella estimoitujen mallien hyvyyttä; tähän käytetään mm. niin sanottuja todennäköisyys- ja kvantiilikuvaajia (probability / PP-plot ja quantile / QQ-plot). Käsiteltäviä menetelmiä sovelletaan reaalimaailman dataan: esimerkkinä tässä luvussa toimii merenpinnan korkeuden mallinnus, jota tarkastellaan yksityiskohtaisesti eri menetelmien yhteydessä. Aloitetaan käytettävän vedenkorkeusdatan kuvauksella. Tämä luo pohjan myöhemmille esimerkeille.

2.1 Johdanto: Merenpinnan korkeus Helsingissä

Suomen rannikolla on mitattu vedenkorkeutta vuodesta 1841 lähtien erilaisilla asteikoilta luettavilla laitteilla, ja mareografeilla eli mittausasemilla vuodesta 1887. Nykyisin mittaukset tehdään automaattisilla laitteistoilla 13 mareografiasemalla.[2]

Itämeren vedenpinnan korkeuteen Suomen etelärannikolla vaikuttavat monet tekijät, joista osa on vaikutuksiltaan pitkäaikaisia ja osa lyhytaikaisia. Merkittävimmiksi tulvavedenkorkeuksiin vaikuttaviksi tekijöiksi on arvioitu Itämeren kokonaisvesimäärä ja ominaisheilahtelu sekä tuuli ja ilmanpaineen vaihtelut. Näistä tuuli ja ilmanpaine ovat merkittävimpiä lyhytaikaisia merivesitulvia aiheuttavia tekijöitä. Myös jääolosuhteet vaikuttavat talvella paikallisiin vedenkorkeuden vaihteluihin. Vuoroveden merkitys Itämeressä sen sijaan on vähäinen. [27, Liite 1]

Tarkastellaan seuraavassa meriveden korkeutta puhtaasti tilastollisena ilmiönä havaintoihin perustuen. Tavoitteena on kuvata poikkeuksellisen korkeiden meriveden tasojen – tai tulvien – sattumista, ja arvioida tähän saakka havaittuja suurempien vedenkorkeuksien sattumisen todennäköisyyttä ääriarvoteorian keinoin. Ongelmanasettelu johtaa siis vedenkorkeuden todennäköisyysjakauman häntätodennäköisyyksien estimointiin.

On hyvä pitää koko ajan mielessä, että kaikki tarkastelu esimerkissä perustuu vedenkorkeuden havaittuihin maksimeihin, ja esimerkiksi vedenpinnan keskimääräistä tasoa ei siis tässä tarkastella. Tulviin varautumisen ja (vahinko)vakuutuksen näkökulmasta vain poikkeuksellisen korkeat, ongelmia aiheuttavat vedenpinnan tasot ovat kiinnostavia. Analyysi perustuu Ilmatieteenlaitokselta

[1] Tullaan näkemään, että stationaariseen prosessiin, jossa havainnot yleisesti ovat riippuvia (serially dependent) voidaan käytännössä soveltaa samoja menetelmiä kuin iid dataankin. Mikäli prosessin ääriarvoilla on taipumusta sattua ryppäissä, datan käsitteleminen kuin iid on periaatteessa ongelmallista. Helpoin lähestymistapa on jättää ongelma huomiotta, ja tulkita suurimman uskottavuuden menetelmä kvasi-SU-menetelmäksi (quasi-maximum likelihood, QML), missä uskottavuusfunktio on väärin määritelty datan riippuvuusrakenteen suhteen. Piste-estimaattien tulisi tästä huolimatta olla kohtuullisen tarkkoja, vaikka näin saadut keskivirheet todennäköisesti ovat liian pieniä. [5] Ks. myös luku 4.

[2] www.itämeriportaali.fi – Vedenkorkeus (`http://www.itameriportaali.fi/fi/tietoa/sanakirja/fi_FI/vedenkorkeus/`)

saatuun aineistoon, joka sisältää vedenkorkeuden päiväkohtaiset maksimiarvot ajalta 1.1.1904–31.12.2011 mitattuna Helsingin mareografilla eli mittausasemalla. Vuosien 1904–1970 aineisto perustuu 6 havaintoon per vuorokausi, ja vuosien 1970–2011 aineisto tasatuntihavaintoihin. Korkeusjärjestelmänä on teoreettinen keskivesi, joka on käytännön tarpeita varten tehty vuosittain muuttuva ennuste vedenkorkeuden pitkäaikaiselle keskiarvolle (odotusarvolle). Teoreettisen keskiveden laskennassa otetaan huomioon maankohoaminen ja merenpinnan nousu.[3]

Kun aineiston vedenkorkeusmaksimit on ilmoitettu (kunkin mittaushetken) teoreettisen keskiveden suhteen[4], on aineistosta siis periaatteessa puhdistettu *keskimääräisen* vedenkorkeuden muutoksen vaikutus vedenkorkeuden maksimeihin, olettaen että keskivesi on määritetty oikein. Tällöin mahdollisten vedenkorkeusmaksimeissa havaittujen trendien tulisi vastata aitoja, nimenomaan vedenkorkeuden ääriarvoihin (tässä maksimeihin) vaikuttavia tekijöitä. Teoreettisen keskiveden määrittämistä ja siihen liittyviä kysymyksiä ei tässä yhteydessä käsitellä sen enempää, vaan data otetaan sellaisenaan analyysin pohjaksi.

Pitkän aikavälin ennusteissa merenpinnan keskimääräisen tason muutos täytyy luonnollisesti huomioida myös maksimien absoluuttisia arvoja muuttavina tekijöinä, tai vastaavasti teoreettisen keskiveden muutos täytyy huomioida kun ennusteen pohjana käytetyt historialliset maksimit on ilmoitettu sen suhteen. Valtamerien pinnan nousu nostaa vedenkorkeutta, kun taas Itämeren alueella yhä jatkuva jääkauden jälkeinen maankohoaminen laskee sitä. Ilmastonmuutoksen voi olettaa vaikuttavan myös merenpinnan korkeuden vaihteluun ja korkeuden ääriarvoihin.

Tarkastellaan seuraavaksi analyysin perustana olevaa vedenkorkeusdataa. Taulukossa 2.1 alla on esitetty joitakin tavanomaisia tilastollisia tunnuslukuja havaintoaineistosta (IQR eli Inter-Quartile Range taulukon viimeisessä sarakkeessa tarkoittaa väliä, jolle mahtuu datan kaksi keskimmäistä kvartiilia, eli jakauman 25 % ja 75 % kvantiilien väliin jäävät havainnot).

Taulukko 2.1: Tilastollisia tunnuslukuja päivittäisten vedenkorkeusmaksimien muodostamalle aikasarjalle.

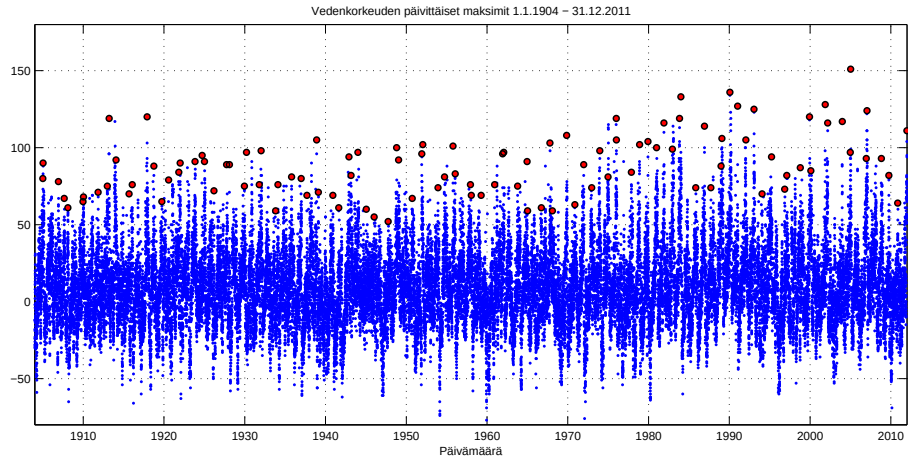
n	min	max	mediaani	moodi	keskiarvo	keskihajonta	IQR
39 447	-77	151	8	10	9.5	24.9	31.0

Vedenkorkeusdata eli päivätason maksimihavainnot on esitetty kuvassa 2.2; vuosittaiset maksimit on merkitty kuvaan punaisilla palloilla. Aineiston korkein mitattu vedenkorkeus oli 151 cm 9.1.2005.

Vedenkorkeusaikasarjaa tarkastellessa havaitaan välittömästi sarjassa ilmenevä kausivaihtelu, siten että vedenkorkeuden päivämaksimit ovat yleisesti ottaen suurimpia talvikuukausina, ja jäävät selvästi pienemmäksi keväällä ja myös kesällä. Tämä näkyy kuvasta paremmin tarkastelemalla hieman lyhyempää ajanjaksoa: kuvassa 2.3 näkyy aikaväli 1.1.2002–31.12.2011. Kuvassa 2.4 on puoles-

[3] www.ilmatieteenlaitos.fi – Mareografi (`http://ilmatieteenlaitos.fi/mareografi`)

[4] Nollatasoksi on siis otettu teoreettinen keskivesi, mistä johtuen aineistossa on myös negatiivisia päivämaksimiarvoja.



Kuva 2.2: Meriveden päivittäiset korkeusmaksimit.

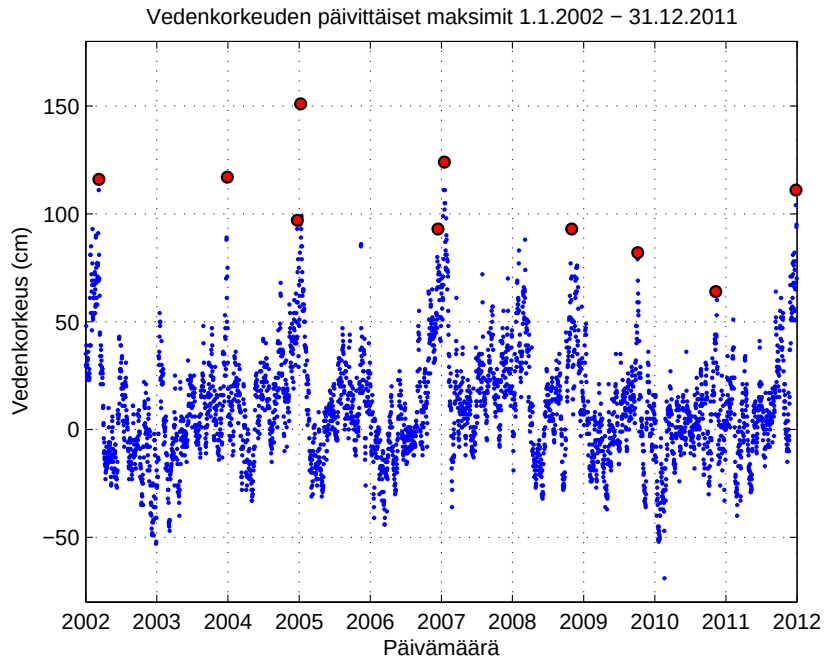
taan esitetty päivämaksimien arvot kuukausittain jaoteltuna. Merenpinnan korkeus käyttäytyy eri tavoin eri vuodenaikoina, mikä johtuu pääasiassa tuulen ja ilmanpaineen käyttäytymisen vuotuisesta kierrosta – yksittäiset vuodet voivat kuitenkin poiketa suuresti toisistaan eikä tällaista keskimääräistä vuodenaikaiskiertoa havaita joka vuosi. Keskimääräinen merenpinnan korkeus on Suomen rannikolla korkeimmillaan joulukuussa ja matalimmillaan huhti-toukokuussa.[5] Tämä heijastuu myös tarkasteltaviin maksimiarvoihin.

Edelleen datasta havaitaan, että päivämaksimit eivät ole riippumattomia, vaan niillä on taipumus klusteroitua eli kerääntyä yhteen (vrt. kuva 2.3): meriveden ollessa korkealla tiettynä päivänä, lisää tämä todennäköisyyttä että myös seuraavana päivänä havaitaan korkea meriveden taso (ts. aikasarja on autokorreloitunut). Tämä on luonnollista ilmiön taustalla olevaa fysikaalista prosessia ajatellen: Itämeren kokonaisvesimäärää säätelee pääasiassa veden virtaus sisään ja ulos suhteellisen matalien ja kapeiden Tanskan salmien läpi. Tulvia aiheuttavien vesimassojen päästyä Itämereen (esim. pitkäaikaisen tuulen vaikutuksesta), vaaditaan aikaa ennen kuin vedenkorkeus laskee takaisin tulvaa edeltäneelle tasolle. Itämeren kokonaisvesimäärän vaihtelusta johtuvat korkeat vedenkorkeudet saattavat kestää viikkoja, jopa kuukausia. [27]

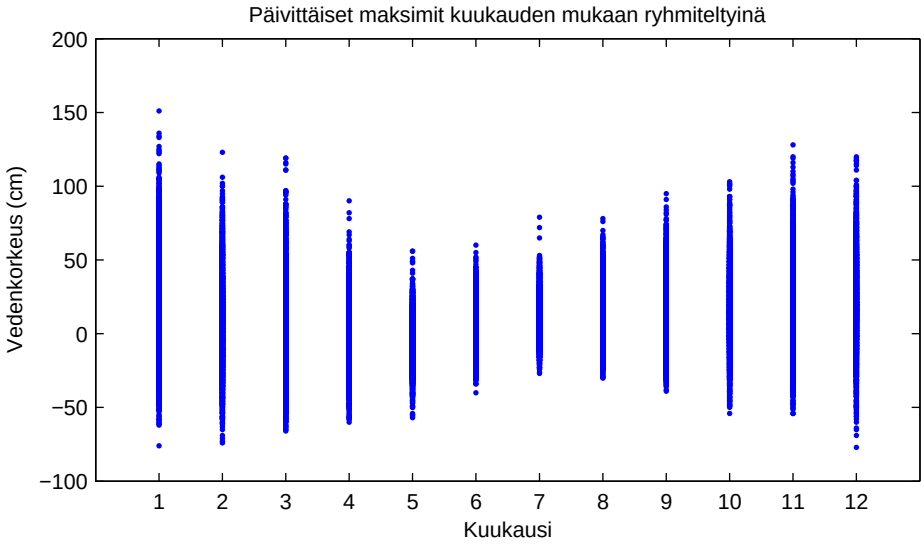
Muilta osin päivämaksimien aikasarja vaikuttaa satunnaiselta, ja siten vaikuttaa perustellulta ottaa aluksi työhypoteesiksi oletus, että datan generoinut stokastinen prosessi on stationaarinen.[6] Tätä oletusta tullaan myöhemmissä osioissa löysentämään.

[5] www.itämeriportaali.fi – Vedenkorkeus (`http://www.itameriportaali.fi/fi/tietoa/yleiskuvaus/veden_liikkeet/vedenkorkeus/`)

[6] Vrt. myös vuosimaksimeja koskevaan tarkasteluun osiossa 2.2.4.



Kuva 2.3: Vedenkorkeuden maksimit aineiston viimeiseltä 10 vuodelta.



Kuva 2.4: Päivämaksimien korkeudet kuukausittain.

2.2 Blokkimaksimimenetelmä

Tarkastellaan dataa, joka muodostuu iid satunnaismuuttujien $X_1, \ldots, X_n$ realisaatiosta, missä satunnaismuuttujien yhteisenä kertymäfunktiona on F. Osion 1.3 teoria antaa otosmaksimin M_n jakaumalle approksimaation

$$\mathbb{P}(M_n \leq x) = F^n(c_n x + d_n) \approx H_\xi(x)$$

suurilla n, kun $F \in \mathrm{MDA}(H_\xi)$. Tämä antaa aiheen mallintaa maksimeja keräämällä lukuisa määrä havaintoja n-blokin otosmaksimeista M_n ja sovittamalla yleistetty ääriarvojakauma (GEV) näihin. Muistetaan GEV-jakauman määritelmä:

$$H_{\xi,\mu,\sigma}(x) = \exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\}, \quad 1+\xi\frac{x-\mu}{\sigma} > 0. \tag{2.1}$$

Tapaus $\xi = 0$ vastaten Gumbel-jakaumaa tulkitaan raja-arvona, kun $\xi \to 0$, ja on siis

$$H_{0,\mu,\sigma}(x) = \exp\left\{e^{-\frac{x-\mu}{\sigma}}\right\}, \quad x \in \mathbb{R}.$$

Merkitään GEV-jakauman parametrijoukkoa lyhyesti $\boldsymbol{\theta} = (\xi, \mu, \sigma) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}_+$, missä tutusti ξ on jakauman muotoparametri, μ lokaatioparametri ja σ skaalaparametri. Jatkossa merkitään tilanteesta riippuen yleistettyä ääriarvojakaumaa joko H_ξ tai $H_{\boldsymbol{\theta}}$.

Huomautettakoon tässä yhteydessä, että käsiteltävä lähestymistapa eli GEV-jakauman sovittaminen otosmaksimeihin voidaan tulkita siten, että implisiittisesti oletetaan havaintojen olevan täsmälleen GEV-jakautuneita, eli $X_1, \ldots, X_n$ iid, $X_i \sim H_{\boldsymbol{\theta}}$, $\forall i$.[7] Tarkkaan ottaen oletus täsmällisestä GEV-jakautuneisuudesta ei ehkä ole realistisin mahdollinen, mutta se edistää mallinnusta suuresti mahdollistamalla parametristen estimointimenetelmien käytön.

Blokkimaksimimenetelmän soveltamiseksi havainnot tulee jakaa samankokoisiin ryhmiin eli blokkeihin. Menetelmän käyttö on siten luonnollisinta, kun on olemassa luonnollinen tapa ryhmitellä havainnot. Esimerkiksi päivittäiset vedenkorkeusmittaukset tai osaketuotot (tappiot) voitaisiin jakaa vuosittaisiin ryhmiin ja ottaa kustakin ryhmästä maksimihavainto.

Oletetaan, että havaintoaineisto on jaettu m:ään n:n havainnon blokkiin, ja merkitään i. blokkia $\boldsymbol{X}^{(i)}$. Tällöin data $(X_i)_{i=1}^{mn}$ voidaan esittää muodossa

$$\begin{aligned} \boldsymbol{X}^{(1)} &= \left(X_1^{(1)}, \ldots, X_n^{(1)}\right) \\ &\vdots \\ \boldsymbol{X}^{(m)} &= \left(X_1^{(m)}, \ldots, X_n^{(m)}\right) \end{aligned}$$

[7] Vaihtoehto parametrisen GEV-perheen sovittamiselle dataan asymptoottisten argumenttien perusteella on tarkastella ns. semiparametrisia menetelmiä, joissa oletetetaan vain, että havainnot $X_i \sim F$, missä $F \in \mathrm{MDA}(H_\xi)$. Nämä keskittyvät muotoparametrin ξ (tai sen käänteisluvun, häntäindeksin $1/\alpha$) estimointiin: tunnettuja esimerkkejä menetelmistä ovat Hill-estimaattori, Pickands-estimaattori, ja Dekkers-Einmahl-de Haan–estimaattori. Semiparametristen menetelmien hyöty käytännön mallinnussovelluksissa on kuitenkin usein rajattu, eikä niiden taustaa käsitellä tarkemmin tässä esityksessä; ks. kuitenkin osio 3.2, jossa esitetään lyhyt käytännön esimerkki. Menetelmistä lisää, ks. [2, kappale 6.4].

Vektoreiden $(\boldsymbol{X}^{(i)})$ oletetaan olevan riippumattomia ja samoin jakautuneita, mutta kunkin vektorin sisällä komponenttien $X_j^{(i)}$ kesken voi olla – ja todellisuudessa yleensä onkin – riippuvuutta. Kustakin havaintovektorista eli n-blokista valitaan jakauman sovitusta varten otosten maksimit, eli

$$M_i = M_{n,i} = \max\left(X_1^{(i)}, \ldots, X_n^{(i)}\right), \quad i = 1, \ldots, m.$$

Oletetaan jatkossa tässä osiossa, että maksimien poimiminen lähtödatasta on jo tehty, ja tarkasteltava data koostuu valmiiksi n-blokkien maksimeista; merkitään näitä mukavuuden vuoksi $M_i = X_i$.

Havaintojen lukumäärä n eli havaintojakson pituus tulisi valita niin, että riippumattomuusoletusta voi pitää perusteltuna. Toisaalta otoksien lukumäärän m tulisi olla riittävän suuri, jotta GEV-jakauman parametrit voidaan luotettavasti estimoida datasta. Blokkien koon n ja lukumäärän m valitsemisessa joudutaankin välttämättä tekemään kompromissi: Suurella n GEV-jakaumaapproksimaatio otosmaksimeille on tarkempi johtaen pienempään harhaan parametriestimaateissa. Suurella m puolestaan on enemmän dataa johon jakauma sovittaa, ja estimoitujen parametrien varianssi on pienempi.

Hyvin usein sovelluksissa n valitaan pragmaattisesti vastaamaan vuoden mittaista periodia. Tämä osoittautuu usein käyttökelpoiseksi valinnaksi, ja kiertää mm. vuoden sisäisistä kausivaihteluista ja sykleistä johtuvat riippuvuusongelmat.

2.2.1 Suurimman uskottavuuden menetelmä GEV-jakaumalle

Suurimman uskottavuuden menetelmä on käytännössä useimmin käytetty parametrinen estimointimenetelmä ääriarvojen tapauksessa, mm. sen hyvien asymptoottisten ominaisuuksien ja erityisesti joustavuuden ansiosta. Liitteessä A on esitetty tarkemmin suurimman uskottavuuden menetelmän taustaa.

Oletetaan, että $\boldsymbol{X} = (X_1, \ldots, X_m)$ missä X_i ovat iid GEV-jakautuneita. Havaittuun otokseen $\boldsymbol{X} = \boldsymbol{x} = (x_1, \ldots, x_m)$ perustuva *uskottavuusfunktio* on tällöin

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = \prod_{i=1}^{m} h_{\boldsymbol{\theta}}(x_i) \mathbb{1}_{\{1+\xi(x_i-\mu)/\sigma>0\}},$$

missä $h_{\boldsymbol{\theta}}$ on GEV-jakauman $H_{\boldsymbol{\theta}}$ tiheysfunktio. Laskennassa on yleensä mukavampi käyttää *logaritmista uskottavuusfunktiota* $l(\boldsymbol{\theta}; \boldsymbol{x}) = \ln L(\boldsymbol{\theta}; \boldsymbol{x})$. GEV-jakauman log-uskottavuudeksi, kun $\xi \neq 0$, saadaan suoralla laskulla

$$\begin{aligned} l(\boldsymbol{\theta}; \boldsymbol{x}) = -m \ln \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \ln\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right] \\ - \sum_{i=1}^{m} \left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]^{-1/\xi}, \end{aligned} \tag{2.2}$$

kun

$$1 + \xi \frac{x_i - \mu}{\sigma} > 0, \quad i = 1, \ldots, m, \tag{2.3}$$

ja $\sigma > 0$. Niillä parametrikombinaatioilla, joilla ehto (2.3) ei täyty – eli kun yksi tai useampi datapiste osuu jakauman määrittelyalueen ulkopuolelle – uskottavuus $L = 0$, ja log-uskottavuus $l = -\infty$. Täten maksimointia ei tarvitse erikseen käsitellä rajoitettuna optimointitehtävänä, kunhan alkuarvot on valittu jotakuinkin järkevästi. Ehdon $\sigma > 0$ täyttyminen on helpointa varmistaa uudelleenparametrisoinnilla asettamalla $\tilde{\sigma} := \ln \sigma$ numeerisessa laskennassa; $\tilde{\sigma}$:n ja σ:n välillä on 1-1–vastaavuus, ja $\tilde{\sigma}$ on aina positiivinen.

Tapauksessa $\xi = 0$ log-uskottavuudeksi tulee GEV-jakauman Gumbel-muodosta lähtien

$$l(\boldsymbol{\theta}; \boldsymbol{x}) = -m \ln \sigma - \sum_{i=1}^{m} \left(\frac{x_i - \mu}{\sigma} \right) - \sum_{i=1}^{m} \exp \left\{ - \left(\frac{x_i - \mu}{\sigma} \right) \right\}, \tag{2.4}$$

kun $\sigma > 0$.

Parametrijoukon $\boldsymbol{\theta} = (\xi, \mu, \sigma)$ suurimman uskottavuuden estimaattori $\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\mu}, \hat{\sigma})$ saadaan maksimoimalla (log-)uskottavuusfunktion arvo. GEV-jakauman tapauksessa eksplisiittistä ratkaisua uskottavuusyhtälöille ei ole olemassa, vaan maksimoinnissa täytyy turvautua numeerisiin menetelmiin.

Mahdollinen ongelma suurimman uskottavuuden menetelmän soveltamisessa ääriarvoteorian jakaumiin kohdataan SU-menetelmään liittyviä säännöllisyysehtoja tarkastellessa. Niin sanotuissa säännöllisissä tapauksissa SU-menetelmän tuottamat estimaatit ovat tehokkaita, tarkentuvia ja asymptoottisesti normaalijakautuneita (ks. esim. [28]). GEV- ja GP-jakaumien kohdalla estimointiongelma on kuitenkin epäsäännöllinen, koska jakaumien määrittelyalue riippuu tuntemattomista, estimoitavista parametreista: $\mu - \sigma/\xi$ on GEV-jakauman vasen päätepiste, kun $\xi > 0$, ja oikea päätepiste, kun $\xi < 0$.

Säännöllisyysehtojen toteutumatta jäämisestä johtuen suurimman uskottavuuden menetelmän standardit asymptoottiset ominaisuudet eivät täten automaattisesti päde, vaikka piste-estimaatti $\hat{\boldsymbol{\theta}}$ onkin mahdollista löytää numeerisesti. Smith [29] osoitti, että SU-estimaattoreiden klassiset (hyvät) ominaisuudet pätevät, kun $\xi > -1/2$. Tarkemmin,

- kun $\xi > -1/2$, SU-estimaattoreilla on niiden tavanomaiset asymptoottiset ominaisuudet;
- kun $-1 < \xi \leq -1/2$, SU-estimaattorit ovat yleisesti saatavissa, mutta standardit asymptoottiset ominaisuudet eivät päde; ja
- kun $\xi < -1$, SU-estimaattorit eivät yleensä ole saatavissa.

Tapaus $\xi < -1/2$ (vastaten erittäin lyhyttä oikeaa häntää) tulee käytännössä hyvin harvoin vastaan; esimerkiksi vakuutus- ja finanssisovelluksissa kohdatuilla jakaumilla on yleensä rajoittamaton oikea häntä, vastaten muotoparametrin arvoa $\xi \geq 0$. Täten suurimman uskottavuuden menetelmän soveltamiselle ei yleensä ole esteitä.

2.2.1.1 Parametriestimaattien luottamusvälit

Asymptoottiseen varianssiin perustuvat luottamusvälit. Suurimman uskottavuuden estimaattoreiden asymptoottisten ominaisuuksien nojalla estimaa-

tit ovat asymptoottisesti normaaleja (tiettyjen säännöllisyysehtojen ollessa voimassa), eli SU-estimaattorille $\hat{\boldsymbol{\theta}}$ pätee

$$\hat{\boldsymbol{\theta}} \dot{\sim} \mathrm{N}_d(\boldsymbol{\theta}, \boldsymbol{I}_E(\boldsymbol{\theta})^{-1}),$$

missä N_d tarkoittaa d-ulotteista multinormaalijakaumaa, d on parametrivektorin $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ dimensio, ja $\boldsymbol{I}_E(\boldsymbol{\theta})$ ns. odotusarvoinen (Fisherin) informaatiomatriisi (ks. liite A). Parametrien θ_i asymptoottinen kovarianssimatriisi on siis informaatiomatriisin $\boldsymbol{I}_E$ käänteismatriisi, ja parametrien varianssit saadaan ko. käänteismatriisin diagonaalilta.

Koska $\boldsymbol{\theta}$:n todellinen arvo on yleensä tuntematon, informaatiomatriisia $\boldsymbol{I}_E$ approksimoidaan usein käyttämällä *havaittua* informaatiomatriisia $\boldsymbol{I}_O(\boldsymbol{\theta})$ evaluoituna pisteessä $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Merkitään käänteismatriisin $\boldsymbol{I}_O(\hat{\boldsymbol{\theta}})^{-1}$ termejä $v_{i,j}$, $i, j = 1, \ldots, d$. Tällöin parametrin θ_i likiarvoiseksi luottamusväliksi luottamustasolla $(1 - \alpha)$ saadaan

$$\left[\hat{\theta}_i - z_{\alpha/2}\sqrt{v_{i,i}},\ \ \hat{\theta}_i + z_{\alpha/2}\sqrt{v_{i,i}}\right],$$

missä $z_{\alpha/2}$ on standardinormaalijakauman $(1 - \alpha/2)$-kvantiili, $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$. Tarkemmat yksityiskohdat löytyvät liitteestä A.

Profiiliuskottavuuteen perustuvat luottamusvälit. Vaihtoehtoinen – ja yleensä tarkempi – menetelmä luottamusvälien rakentamiseen perustuu ns. profiiliuskottavuuteen (profile likelihood). Log-uskottavuus voidaan formaalisti kirjoittaa $l(\boldsymbol{\theta}) = l(\theta_i, \boldsymbol{\theta}_{-i})$, missä $\boldsymbol{\theta}_{-i}$ tarkoittaa kaikkien muiden kuin i. komponentin muodostamaa vektoria. Parametrin θ_i *profiiliuskottavuus* (tai tässä *profiililog-uskottavuus*) määritellään

$$l_p(\theta_i) = \max_{\boldsymbol{\theta}_{-i}} l(\theta_i, \boldsymbol{\theta}_{-i})$$

Siis kullekin θ_i:n arvolle profiiliuskottavuus on kaikkien muiden parametrien suhteen maksimoitu log-uskottavuus. Voidaan myös ajatella, että $l_p(\theta_i)$ on log-uskottavuuspinnan profiili parametriavaruuden θ_i-akselilta nähtynä; tästä nimi "profiiliuskottavuus".

Profiiliuskottavuus parametreille $\boldsymbol{\theta} = (\xi, \mu, \sigma)$ saadaan siis suoraviivaisesti maksimoimalla log-uskottavuusfunktio muiden parametrien suhteen. Esimerkiksi jakuman muotoparametria ξ tarkastellessa kiinnitetään ensin $\xi = \xi_0$, ja maksimoidaan log-uskottavuus (2.2) (tai vastaava Gumbel-versio (2.4)) parametrien μ ja σ suhteen. Toistamalla tämä useilla eri arvoilla ξ_0 saadaan rakennettua profiiliuskottavuus parametrille ξ. Likimääräiset luottamusvälit voidaan konstruoida uskottavuusosamäärätestiin (likelihood ratio test) perustuen; ks. liite B.

2.2.2 Toistumisperiodi ja toistumistaso

Esitetään tässä vaiheessa toistumisperiodin ja toistumistason määritelmät. Nämä käsitteet ovat hyödyllisiä yleisesti ääri-ilmiöitä hahmotettaessa ja erityisesti estimoitujen mallien antamia tuloksia tulkittaessa.[8] *Toistumisperiodi* (tai *toistumisjakso*) vastaa kysymykseen "mikä on keskimääräinen odotusaika kahden

[8] Tosin hyöty rajoittuu pitkälti stationaaristen mallien tarkasteluun, kuten hieman myöhemmin tullaan näkemään.

määrätyn, samantyyppisen ääritapahtuman välillä?", ja t. vuoden *toistumistaso* vastaavasti kysymykseen "mikä on se määrätyn ääri-ilmiön taso, joka ylitetään keskimäärin kerran t:ssä vuodessa?".

Käsitteet ovat intuitiivisesti selkeitä, mutta esitetään ensin hieman taustaa ennen muodollista määritelmää. Olkoon (X_i) jono iid satunnaismuuttujia jatkuvalla kertymäfunktiolla F, ja olkoon u kiinnitetty kynnystaso. Tarkastellaan tason u ylitteiden jonoa $(\mathbb{1}_{\{X_i>u\}})$; nämä ovat iid Bernoulli-jakautuneita satunnaismuuttujia onnistumistodennäköisyydellä $p = \bar{F}(u)$. Tästä seuraa, että ensimmäisen "onnistumisen" eli ensimmäisen tason u ylitteen sattumisaika,

$$L(u) = \min_{i\geq 1}\{X_i > u\},$$

on geometrisesti jakautunut satunnaismuuttuja jakaumalla

$$\mathbb{P}(L(u) = k) = (1-p)^{k-1}p, \quad k = 1, 2, \ldots.$$

Edelleen satunnaismuuttujat

$$L_n(u) = \min_{i>L_{n-1}} \{X_i > u\} - L_{n-1}(u), \quad n > 1,$$

kuvaavat ajanjaksoja kahden peräkkäisen tason u ylityksen välillä (asetetaan $L_1(u) = L(u)$). Ylitysajat ovat riipumattomia, koska X_i:t ovat. Tapahtumien $\{X_i > u\}$ toistumisperiodi on nyt $\mathbb{E}(L(u)) = 1/p = 1/\bar{F}(u)$.

Määritelmä 2.1 (Toistumisperiodi)
Olkoon (X_i) jono satunnaismuuttujia yhteisenä kertymäfunktiona F. Tapahtumien $\{X_i > u\}$ toistumisperiodi (return period) *on*

$$t_u = \frac{1}{\bar{F}(u)}.$$

Toistumisperiodin tulkinnan mukaan ajassa t_u (t_u:ssa havainnossa) sattuu keskimäärin yksi tapahtuma (tason u ylitys), tai täsmällisemmin, aikayksikössä sattuu tapahtuma todennäköisyydellä $\bar{F}(u)^{-1}$. Toistumisperiodin (ja toistumistason) tulkinta on erityisen selkeä, kun aikayksikkö vastaa yhtä vuotta, kuten asianlaita usein blokkimaksimimenetelmässä on. Tällöin on luonnollista puhua odotusperiodin yhteydessä "t. vuoden tapahtumasta". Mikäli kuitenki ääriarvoilla on taipumus kasaantua tarkateltavassa prosessissa riippuvuuden seurauksena, useamman tapahtuman voidaan odottaa sattuvan lähekkäin (samana vuonna), vaikka keskimäärin yksi tapahtuma sattuukin t_u vuodessa.

Määritelmä 2.2 (Toistumistaso)
Olkoon (X_i) jono satunnaismuuttujia yhteisenä kertymäfunktiona F. (Toistumis)periodia t vastaava toistumistaso (return level) *on*

$$x_t = q_{1-1/t}(F) = F^{\leftarrow}\left(1 - \frac{1}{t}\right).$$

Toistumistaso on siis jakauman $(1-1/t)$-kvantiili, ja se voidaan tulkita tasoksi, joka ylitetään keskimäärin kerran ajassa t (tai t:ssä havainnossa). Määritelmistä

seuraa, että periodin (vuoden, havainnon) t_u toistumistaso on u. Epästationaarisessa tapauksessa toistumisperiodin ja toistumistason määrittely ei ole yksikäsitteistä, koska havaintojen jakauma ja siten tapahtumien todennäköisyydet muuttuvat ajassa. Tähän palataan epästationaaristen prosessien tarkastelun yhteydessä osiossa 2.5.

Merkitään toistumisperiodia lyhyesti $1/p$ ja tätä vastaavaa toistumistasoa x_p, $0 < p < 1$. p on esimerkiksi 0.01 tai 0.001, vastaten vuosimaksimimallissa tapahtumia, jotka sattuvat keskimäärin kerran 100 tai 1 000 vuodessa. Kvantiilille x_p saadaan eksplisiittinen esitys kääntämällä yhtälö (2.1):

$$x_p = H^{-1}_{\xi,\mu,\sigma}(1-p) = \begin{cases} \mu - \frac{\sigma}{\xi}(1 - \{-\ln(1-p)\}^{-\xi}), & \xi \neq 0, \\ \mu - \sigma \ln\{-\ln(1-p)\}, & \xi = 0. \end{cases} \tag{2.5}$$

Toistumistason estimaatti $\hat{x}_p$ saadaan edelleen sijoittamalla yo. yhtälöön estimoidut parametrit $\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\mu}, \hat{\sigma})$.

Kvantiilien tarkasteleminen tarjoaa kätevän tavan arvioida sovitettua GEV-mallia. Graafista tarkastelua varten on mukava asettaa vielä $y_p = -\ln(1-p)$, jolloin yhtälö (2.5) voidaan kirjoittaa

$$x_p = H^{-1}_{\xi,\mu,\sigma}(1-p) = \begin{cases} \mu - \frac{\sigma}{\xi}(1 - y_p^{-\xi}), & \xi \neq 0, \\ \mu - \sigma \ln y_p, & \xi = 0. \end{cases} \tag{2.6}$$

Kun x_p piirretään y_p:n funktiona logaritmisella skaalalla (tai vaihtoehtoisesti, x_p piirretään $\ln y_p$:n funktiona normaalilla skaalalla), näin saatu graafi – toistumistasokuvaaja, return level plot – on lineaarinen, jos $\xi = 0$. Jos taas $\xi < 0$, kuvaaja on konveksi asymptoottisena rajatasona $\mu - \sigma/\xi$, kun $p \to 0$; ja jos $\xi > 0$, kuvaaja on konkaavi ja kasvaa äärettömiin. Toistumistasokuvaaja on erityisen hyödyllinen mallin validoinnissa ja toisaalta myös esittämisessä, koska skaalan valinta tiivistää jakauman hännän niin että datan ekstrapolaatio korostuu.

Toistumistasokuvaaja koostuu toistumistasosta toistumisperiodin funktiona (sopivalla skaalauksella). Toistumisperiodikuvaaja saadaan tästä yksinkertaisesti kääntämällä akselit toisinpäin. Molemmat kuvaajat siis sisältävät kaiken olennaisen tiedon tältä osin.

2.2.2.1 Luottamusvälit

Piste-estimaattien lisäksi on tarpeen tarkastella estimaatteihin liittyvää epävarmuutta.

Asymptoottiseen varianssiin perustuvat luottamusvälit. Toistumistaso (ja -periodi) on estimoitujen parametrien yksinkertainen funktionaali. Deltamenetelmää käyttämällä (ks. liite A) saadaan estimaatin $\hat{x}_p$ likimääräiseksi varianssiksi

$$\mathrm{Var}(\hat{x}_p) = \nabla x_p^T \boldsymbol{V}_{\hat{\boldsymbol{\theta}}} \nabla x_p,$$

missä $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}$ on parametriestimaattien $\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\mu}, \hat{\sigma})$ (asymptoottinen) kovarianssi-

matriisi, ja gradientti on

$$\nabla x_p = \begin{pmatrix} \frac{\partial x_p}{\partial \xi} \\ \frac{\partial x_p}{\partial \mu} \\ \frac{\partial x_p}{\partial \sigma} \end{pmatrix} = \begin{pmatrix} \frac{\sigma}{\xi^2}\left(1 - y_p^{-\xi}\right) - \frac{\sigma}{\xi} y_p^{-\xi} \ln y_p \\ 1 \\ -\xi^{-1}\left(1 - y_p^{-\xi}\right) \end{pmatrix}$$

evaluoituna pisteessä $(\hat{\xi}, \hat{\mu}, \hat{\sigma})$. Varianssiin perustuen voidaan suoraan määrittää luottamusvälit toistumistasolle.

Mikäli $\xi < 0$, myös jakauman oikealle päätepisteelle $x_F < \infty$ voidaan johtaa luottamusvälit. Oikea päätepiste vastaa tapausta $p = 0$, eli "äärettömän pitkää" toistumisperiodia. Päätepisteen SU-estimaatti on

$$\hat{x}_0 = \hat{x}_F = \hat{\mu} - \hat{\sigma}/\hat{\xi},$$

ja varianssi saadaan määrättyä kuten aiemminkin, sillä erotuksella, että gradientti on nyt

$$\nabla x_0 = \begin{pmatrix} \sigma\xi^2 \\ 1 \\ -\xi^{-1} \end{pmatrix}$$

evaluoituna pisteessä $(\hat{\xi}, \hat{\mu}, \hat{\sigma})$.

Profiiliuskottavuuteen perustuvat luottamusvälit. Luottamusvälien rakentaminen toistumistasolle x_p profiiliuskottavuuteen perustuen voidaan tehdä parametrisoimalla GEV-malli uudelleen, siten että x_p on yksi malliparametreista. (2.6) antaa suoraan

$$\mu = \begin{cases} x_p + \frac{\sigma}{\xi}(1 - y_p^{-\xi}), & \xi \neq 0, \\ x_p + \sigma \ln y_p, & \xi = 0. \end{cases}$$

ja sijoittamalla tämä log-uskottavuusfunktioon (2.2) saadaan GEV-malli ilmaistua parametrien (ξ, x_p, σ) avulla. Profiiliuskottavuus x_p:lle voidaan nyt määrittää maksimoimalla log-uskottavuus muiden parametrien suhteen,

$$l_p(x_p) = \max_{(\xi,\sigma)} l(x_p; \xi, \sigma),$$

sopivalla välillä $x_p \in \left[x_p^l,\ x_p^r\right]$ luottamustasosta riippuen (ks. liite B).

2.2.3 Mallidiagnostiikkaa

Kaikessa tilastollisessa mallinnuksessa seuraava askel mallin sovittamisen jälkeen on tarkastella estimoidun mallin hyvyyttä. Ääriarvoteoriaan perustuvien mallien kohdalla tilanne on erityisen haastava, sillä koko mallinrakennuksen tavoitteena on yleensä tehdä päätelmiä tähän asti havaittuja äärimmäisempien tapahtumien sattumistodennäköisyydestä, toisin sanoen ekstrapoloida havaitun datan ulkopuolelle.

Vaikka GEV- tai muuhun ääriarvomalliin perustuvan ekstrapolaation oikeellisuutta on mahdotonta todentaa, mallia voidaan ja tulee verrata siihen havaintodataan jota on saatavilla. (Lisäksi usein mallin hyvyyden ja järkevyyden tarkasteluun saadaan tukea sovellusaluekohtaisista fysikaalisista tarkasteluista ja reunaehdoista.) Mallin hyvä sopivuus havaittuun dataan ei tietenkään riitä todistamaan että malli toimisi myös havaintoalueen ulkopuolella, mutta sitä voidaan pitää järkevänä edellytyksenä tälle. Ääriarvoteoriaan perustuvien mallien vahvuutena on se, että matemaattinen teoria kertoo tämän olevan oikea malliluokka maksimien käytöksen mallintamiseen kun otoskoko kasvaa äärettömäksi. Mallintaja voi siis luottaa työskentelevänsä hyvän malliluokan kanssa. Toisaalta teorian tulokset ovat luonteeltaan asymptoottisia, eivätkä päde täsmälleen kuin idealisoiduissa tapauksissa; käytännössä GEV- tai GP-approksimaation pätevyyttä on vaikea arvioida suurillakaan n, ja mallin ekstrapoloiminen perustuu aina todentamattomiin oletuksiin.

Pitäen edellä sanottu mielessä, paras mitä voidaan mallidiagnostiikan mielessä tehdä on verrata havaittuja arvoja mallin antamiin. Esitetään seuraavaksi kaksi graafista työkalua mallin hyvyyden (goodness-of-fit) tarkasteluun, nimittäin todennäköisyyskuvaaja ja kvantiilikuvaaja. Tätä varten määritellään ensin havaintoihin perustuva empiirinen kertymäfunktio (empirical cumulative distribution function, lyh. ECDF).

Määritelmä 2.3 (Empiirinen kertymäfunktio)
Olkoon $X_1, X_2, \ldots$ jono iid satunnaismuuttujia yhteisenä kertymäfunktiona F. Järjestetyn otoksen

$$X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$$

empiirinen kertymäfunktio F_n *määritellään*

$$F_n(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} = \frac{i}{n+1}, \text{ kun } X_{i,n} \leq x < X_{i+1,n}, \quad x \in \mathbb{R}.$$

Huomautus 2.4 *Mille tahansa järjestystunnusluvulle $X_{i,n}$ täsmälleen i otoksen n:stä havainnosta on pienempiä tai yhtä suuria kuin $X_{i,n}$: täten saadaan empiirinen estimaatti $\mathbb{P}(X \leq X_{i,n}) = F_n(X_{i,n}) = i/n$. Tämän sijasta käytetään yleensä muotoa $F_n(X_{i,n}) = i/(n+1)$, jotta vältytään tilanteelta $F_n(X_{n,n}) = 1$ otoksen suurimman havainnon kohdalla.*

Koska empiirinen kertymäfunktio F_n on todellisen havaintojen taustalla olevan todennäköisyysjakauman F estimaatti – itse asiassa harhaton ja tarkentuva sellainen – tulisi F_n:n ja havainnoista estimoidun parametrisen jakaumaestimaatin $\hat{F}$ välillä olla kohtuullinen vastaavuus. Mikäli näin ei ole, estimaattia $\hat{F}$ ei yleensä voi pitää riittävänä mallina havainnoille. Seuraavat diagnostiset menetelmät perustuvat olennaisesti empiirisen kertymäfunktion ja (suurimman uskottavuuden menetelmällä tai jollain muulla menetelmällä) estimoidun kertymäfunktion vertaamiseen.

Todennäköisyyskuvaaja. Aloitetaan määritelmällä.

Määritelmä 2.5 (Todennäköisyyskuvaaja (PP-plot))
Olkoon

$$X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$$

iid havaintojen järjestetty otos, ja $\hat{F}$ populaation estimoitu kertymäfunktio. Todennäköisyyskuvaaja *muodostuu pisteistä*

$$\left\{ \left(\frac{i}{n+1}, \hat{F}(X_{i,n}) \right) : i = 1, \ldots, n \right\}. \tag{2.7}$$

Mikäli $\hat{F}$ on kohtuullinen estimaatti populaatiojakaumalle, todennäköisyyskuvaajan pisteiden tulisi sijoittua lähelle yksikködiagonaalia. Huomattavat poikkeamat lineaarisuudesta puolestaan viittaavat siihen, että malli ei pysty kuvaamaan havaintoja hyvin.

Soveltaen määritelmää GEV-jakaumaan ja havaittuun otokseen ($x_{1,n} \leq \ldots \leq x_{n,n}$), todennäköisyyskuvaajaksi saadaan

$$\left\{ \left(\frac{i}{n+1}, \hat{H}_\xi(x_{i,n}) \right) : i = 1, \ldots, n \right\}, \tag{2.8}$$

missä tutusti

$$\hat{H}_\xi(x) = \exp\left\{ -\left[1 + \hat{\xi}\left(\frac{x - \hat{\mu}}{\hat{\sigma}} \right) \right]^{-1/\hat{\xi}} \right\}.$$

Kvantiilikuvaaja.

Määritelmä 2.6 (Kvantiilikuvaaja (QQ-plot))
Olkoon

$$X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$$

iid havaintojen järjestetty otos, ja $\hat{F}$ populaation estimoitu kertymäfunktio. Kvantiilikuvaaja *muodostuu pisteistä*

$$\left\{ \left(\hat{F}^{\leftarrow}(\frac{i}{n+1}), X_{i,n} \right) : i = 1, \ldots, n \right\}. \tag{2.9}$$

Siinä missä todennäköisyyskuvaaja vertasi empiirisen ja estimoidun jakauman antamia todennäköisyyksiä, kvantiilikuvaaja vertaa nimensä mukaisesti näiden antamia kvantiileita. Mikäli $\hat{F}$ on kohtuullinen estimaatti todelliselle populaatiojakaumalle, kvantiilikuvaajan tulisi jälleen koostua pisteistä jotka ovat lähellä diagonaalia. Vaikka todennäköisyys- ja kvantiilikuvaajat esittävät periaatteessa saman informaation eri skaaloilla, on käytetyllä skaalallakin käytännössä merkitystä. Tästä johtuen kvantiilikuvaaja on todennäköisyyskuvaajaa enemmän käytetty – ja antaa yleensä paremman kuvan mallin hyvyydestä – ääriarvomallien yhteydessä, sillä se tuo jakauman hännän paremmin esiin.

GEV-jakaumalle kvantiilikuvaaja on

$$\left\{ \left(\hat{H}_\xi^{-1}(\frac{i}{n+1}), x_{i,n} \right) : i = 1, \ldots, n \right\}, \tag{2.10}$$

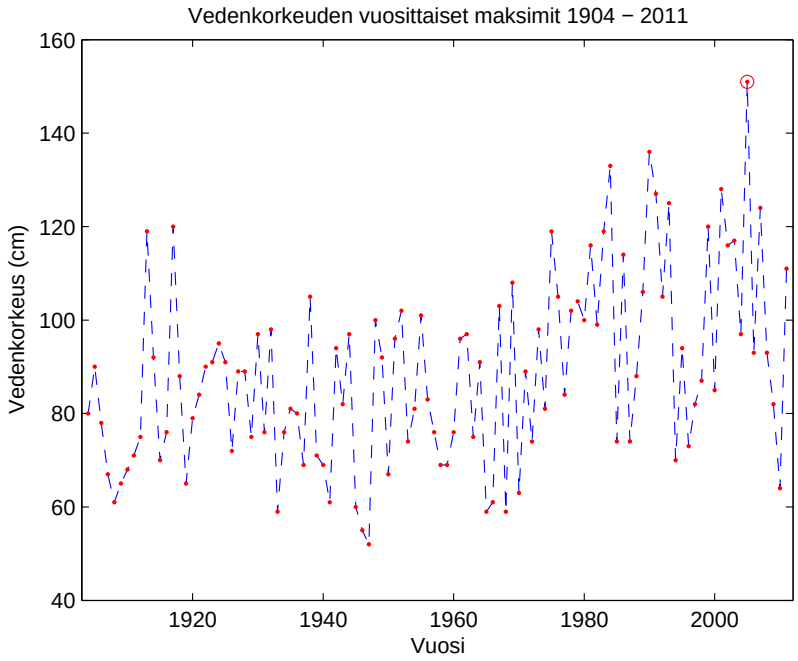
missä

$$\hat{H}_\xi^{-1}(x) = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} \left[1 - \{-\ln x\}^{-\hat{\xi}} \right].$$

Huomautetaan vielä, että mikäli yo. kaavaan sijoitetaan $x = 1 - p$, päädytään kaavaan (2.5) eli toistumisperiodia $1/p$ vastaavaan toistumistasoon riippumattomassa (tai stationaarisessa) mallissa, jälkimmäisen ollessa – ei mitään muuta kuin – jakauman kvantiili.

2.2.4 Merenpinnan korkeuden maksimien mallintaminen GEV-jakaumalla

Tarkastellaan seuraavaksi vuosittaisia vedenkorkeusmaksimeja (kuva 2.5). Kuvan aikasarja on siis saatu poimimalla päivätason datasta kunkin vuoden päivämaksimeista suurin. Näin saadun aikasarjan voidaan olettaa koostuvan (likimain) riippumattomista havainnoista.



Kuva 2.5: Vedenkorkeuden vuosittaiset maksimit.

Taulukossa 2.2 on esitetty joitakin tavanomaisia tunnuslukuja vuosimaksimien sarjalle.

Taulukko 2.2: Tilastollisia tunnuslukuja vedenkorkeushavaintojen vuosimaksimien muodostamalle aikasarjalle.

n	min	max	mediaani	moodi	keskiarvo	keskihajonta	IQR
108	52	151	88	76	88.7	20.2	26.5

Vuosimaksimeja tarkastelemalla kierretään kausivaihtelusta ja syklisyydestä johtuvat ongelmat. Meriveden korkeutta fysikaalisena prosessina ajatellen voidaan esittää intuitiivisia perusteluja myös sille, että vuosittaisten maksimien välillä ei ole merkittävää riippuvuusrakennetta: meriveden korkeus tiettynä päivänä voi vaikuttaa seuraavina päivinä havaittavaan vedenkorkeuteen, mutta tuskin vedenkorkeuteen määrättynä päivänä esimerkiksi vuoden päästä.[9] Tehdään siis

[9]Tarkkaan ottaen vuosittaisten maksimien tarkastelussakin tulee se ongelma, että kahden

oletus, että vedenkorkeus ilmiönä ei ilmennä merkittävää pitkän aikavälin riippuvuutta (long-range dependence) kiinnostuksen kohteena olevilla korkeilla tasoilla; tällöin itse asiassa stationaariseenkin aikasarjaan voidaan soveltaa samoja menetelmiä kuin riippumattomien ja samoin jakautuneiden havaintojen muodostamaan, ks. osiot 1.4 ja 2.4.

Palataan vielä kuvan 2.5 tarkasteluun. Kyseisessä vuosittaiset maksimit sisältävästä aikasarjassa on havaittavissa mahdollinen kasvava trendi. Kuvassa 2.2 (tai 2.3) vastaavaa ei sen sijaan selvästi havaita, kun tarkastellaan koko päivämaksimien aikasarjaa. Näyttää siis mahdolliselta, että vedenkorkeuden vuosittaiset maksimiarvot olisivat keskimäärin kasvaneet pitkän, 108 vuotta kattavan tarkasteluajanjakson aikana. Tähän palataan kappaleissa 2.5 ja 2.6 epästationaarisiin aikasarjoihin soveltuvien menetelmien tarkastelun yhteydessä; toistaiseksi tehdään käsittelyä yksinkertaistava oletus, että data on stationaarista ja mahdollinen trendi jätetään huomiotta.

Edetään sovittamaan GEV-jakauma vuosimaksimidataan. Parametrien estimointi tapahtuu maksimoimalla GEV-jakauman log-uskottavuusfunktio (2.2) numeerisesti MATLABin `fminsearch`-funktiota käyttäen. Tuloksena saadaan parametrien suurimman uskottavuuden estimaatit (SUE) $\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\mu}, \hat{\sigma}) = (-0.0849, 80.0, 17.3)$. Myöhempää vertailua varten log-uskottavuudeksi tulee maksimissa -473.4. Parametriestimaattien likimääräinen kovarianssimatriisi on

$$\boldsymbol{V} = \begin{pmatrix} 0.00634 & -0.0500 & -0.0625 \\ -0.0500 & 1.915 & 0.854 \\ -0.0625 & 0.854 & 3.611 \end{pmatrix},$$

missä matriisin diagonaalilla on siis parametriestimaattien varianssit. Ottamalla näistä neliöjuuret saadaan keskivirheiksi $\mathrm{se}(\hat{\boldsymbol{\theta}}) = \mathrm{se}((\hat{\xi}, \hat{\mu}, \hat{\sigma})) = (0.0797, 1.384, 1.900)$. Likimääräisiksi 95 %:n luottamusväleiksi tulee täten $\hat{\theta}_i \pm z_{\alpha/2}\mathrm{se}(\hat{\theta}_i)$, missä $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2) = 1.960$ kun $1-\alpha = 0.95$ eli $\alpha = 0.05$. Taulukossa 2.3 on esitetty parametriestimaatit luottamusväleineen.

Taulukko 2.3: Parametrien SU-estimaatit luottamusväleineen GEV-mallissa.

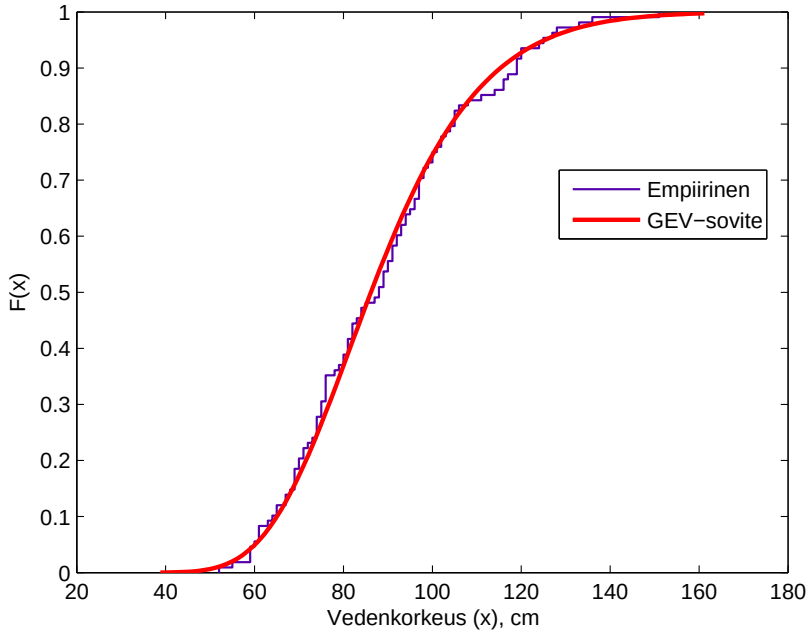
Parametri	**SUE**	**95 %:n luottamusväli**
ξ	-0.0849	$[-0.241, 0.0712]$
μ	80.0	$[76.2, 84.7]$
σ	17.3	$[14.7, 20.2]$
x_F	283	$[-18, 585]$

Muotoparametrin ξ estimaatti on negatiivinen, mikä viittaa jakaumaan jolla on äärellinen oikea päätepiste. Tämä lienee uskottavaa ajateltaessa taustalla olevaa fysikaalista prosessia eli merenpinnan korkeutta. Toisaalta parametrin ξ luottamusväli sisältää arvon 0, eli hypoteesia $\xi = 0$ ei tässä voida hylätä 95 %:n luottamustasolla (5 %:n merkitsevyystasolla).

peräkkäisen vuoden maksimit saattavat olla hyvin lähellä toisiaan, jos vesi pysyy korkealla juuri vuodenvaihteen aikoihin; vrt. vuosien 2004–2005 ja 2006–2007 vuosimaksimeja kuvassa 2.3. Periaatteessa dataan voitaisiin soveltaa sopivaa deklusterointisääntöä, siten että kultakin vuodelta valitun maksimin vaadittaisiin olevan vähintään tietyn etäisyyden (päivien lukumäärän) päässä vierekkäisten vuosien maksimeista.

Päätepisteen estimaatiksi mallissa saadaan $\hat{x}_F = \hat{\mu} - \hat{\sigma}/\hat{\xi} = 283$ (cm); 95 %:n luottamusväli päätepisteelle on laaja, sisältäen alueen negatiivisesta vajaaseen 6 metriin. Nämä on esitetty myös taulukossa 2.3. Luottamusvälin vasemman pisteen negatiivisuus tuo esiin yhden asymptoottiseen kovarianssimatriisiin perustuvien, välttämättä symmetristen luottamusvälien ongelmista. Myöhemmin käsiteltävällä profiiliuskottavuusmenetelmällä saadaankin tarkemmat – ja tässä tapauksessa realistisemmat – luottamusvälit.

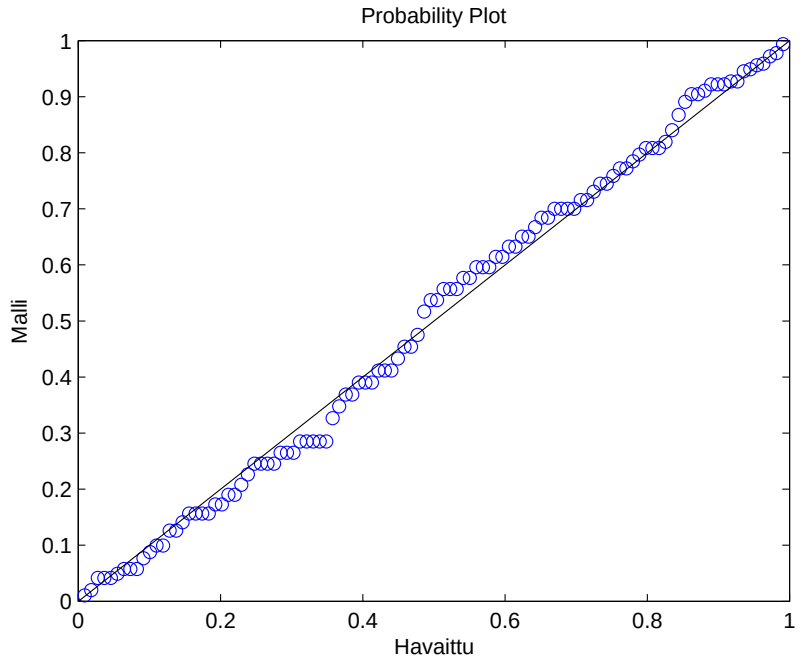
Kuvassa 2.6 on esitetty vedenkorkeusdatan vuosimaksimien empiirinen kertymäfunktio ja estimoitu GEV-jakauman kertymäfunktio. Malli näyttää sopivan havaintoihin varsin hyvin.



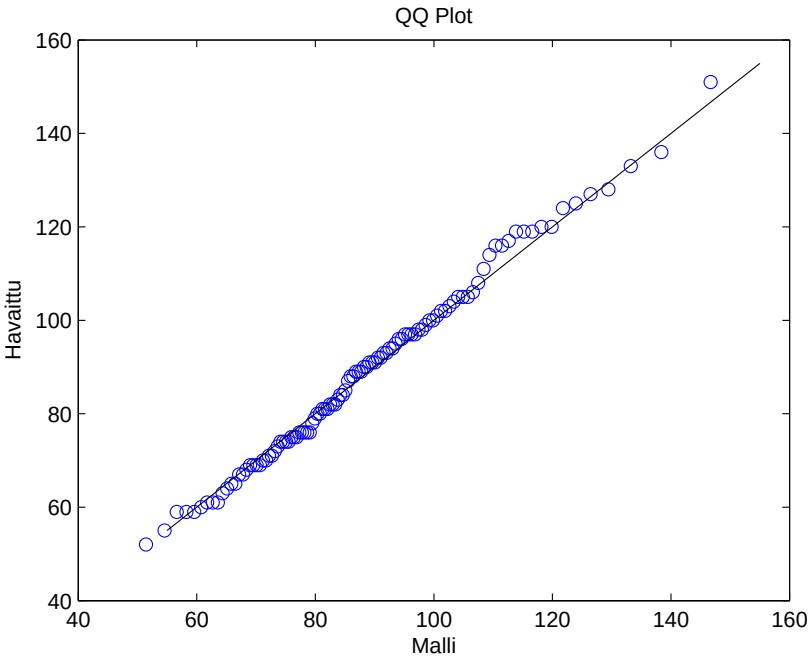
Kuva 2.6: Vedenkorkeuden vuosimaksimien empiirinen kertymäfunktio vs. GEV-sovite.

Muodostetaan sovitetulle mallille osion 2.2.3 todennäköisyys- ja kvantiilikuvaajat (kuva 2.7 ja kuva 2.8). Näiden perusteella malli näyttää sopivan havaintoihin hyvin, kuvat eivät osoita evidenssiä mallia vastaan.

Estimoidusta jakaumasta (kuva 2.6) nähdään suoraan mallin implikoimat merenpinnan eri korkeustasojen ylittämistodennäköisyydet. Olkoon kiinnostuksen kohteena oleva taso x_0; tällöin todennäköisyydellä $H_{\hat{\boldsymbol{\theta}}}(x_0)$ tasoa ei ylitetä vuoden aikana, tai ts. taso ylitetään todennäköisyydellä $\bar{H}_{\hat{\boldsymbol{\theta}}}(x_0) = 1 - H_{\hat{\boldsymbol{\theta}}}(x_0)$. Esimerkiksi korkeinta havaittua vedenkorkeutta, 151 cm, vastaava ylitystodennäköisyys on mallin mukaan $H_{\hat{\boldsymbol{\theta}}}(x_0) = 0.0063$, eli todennäköisyys ylittää 151 cm vuoden aikana on 0.63 %. Vastaavasti mikäli tarkastellaan tiettyä todennäköisyyttä p_0 ja kysytään, mikä on vedenkorkeuden taso jota ei ylitetä tällä todennäköisyydellä vuoden aikana, on vastaus $x_{p_0} = H^{-1}_{\hat{\boldsymbol{\theta}}}(p_0)$. Esimerkik-



Kuva 2.7: Todennäköisyyskuvaaja vuosimaksimien GEV-sovitteelle.



Kuva 2.8: Kvantiilikuvaaja vuosimaksimien GEV-sovitteelle.

si 99.5 %:n todennäköisyydellä tasoa $x_{0.995} = H_{\hat{\boldsymbol{\theta}}}^{-1}(0.995) \approx 154$ cm ei ylitetä vuoden kuluessa; merenpinnan korkeuden yhden vuoden 99.5 % Value-at-Risk (VaR) on siis 154 cm. Taulukkoon 2.4 on kerätty joitakin todennäköisyyksiä vastaavia vedenkorkeustasoja: tulkinta on siis, että todennäköisyydellä p mainittua tasoa $H_{\hat{\boldsymbol{\theta}}}^{-1}(p)$ ei ylitetä vuoden aikana, ja todennäköisyydellä $1-p$ taso ylitetään.

Taulukko 2.4: Eri todennäköisyyksiä vastaavia merenpinnan korkeustasoja GEV-mallissa.

p	$1-p$	$H_{\hat{\theta}}^{-1}(p)$
0.5	0.5	86 (cm)
0.75	0.25	100 (cm)
0.90	0.10	115 (cm)
0.99	0.01	146 (cm)
0.999	0.001	170 (cm)

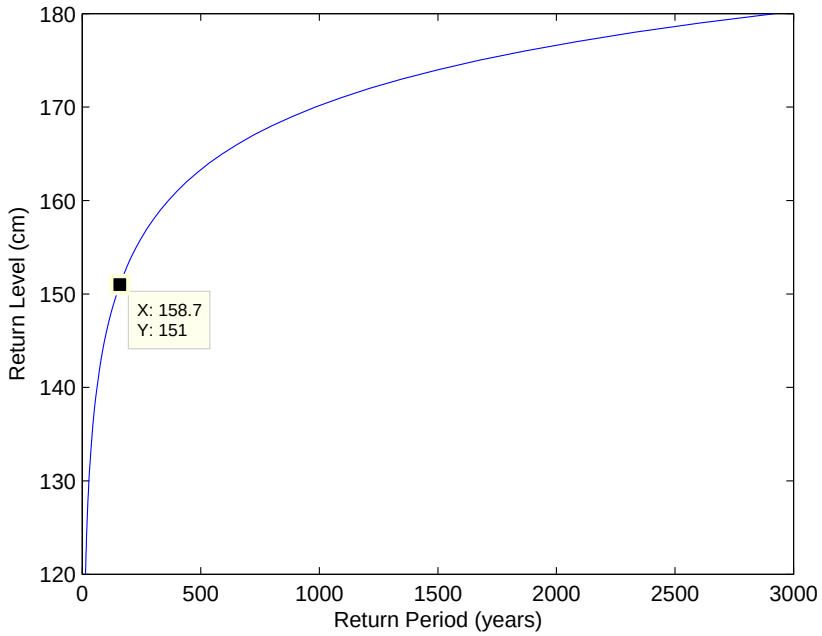
Usein on helpompaa hahmottaa (luonnon)ilmiöitä käyttäen "t. vuoden ilmiö"–terminologiaa. Käytetään sovitettua mallia merenpinnan korkeuden toistumistasojen ja toistumisaikojen estimoimiseen. Kuvassa 2.9 on esitetty toistumistaso toistumisajan funktiona. Toistumisaika toistumistason funktiona puolestaan on vain edellinen käännettynä; hahmottamisen helpottamiseksi tämäkin on esitetty kuvassa 2.10. Kuvaajista nähdään, että esimerkiksi suurin mitattu merenpinnan korkeushavainto 151 cm vastaa mallin mukaan n. 160 vuoden tapahtumaa, eli tapahtumaa, joka sattuu kerran 160 vuodessa (tietysti sillä oletuksella, että ilmiön satunnaisluonne (todennäköisyysjakauma) säilyy ajan suhteen muuttumattomana)[10].

Toistumistasoille (yhtäpitävästi toistumisperiodeille) voidaan rakentaa (likimääräiset) luottamusvälit SU-estimaattien asymptoottiseen kovarianssimatriisiin perustuen delta-menetelmän avulla, osiossa 2.2.2.1 esitetyllä tavalla.

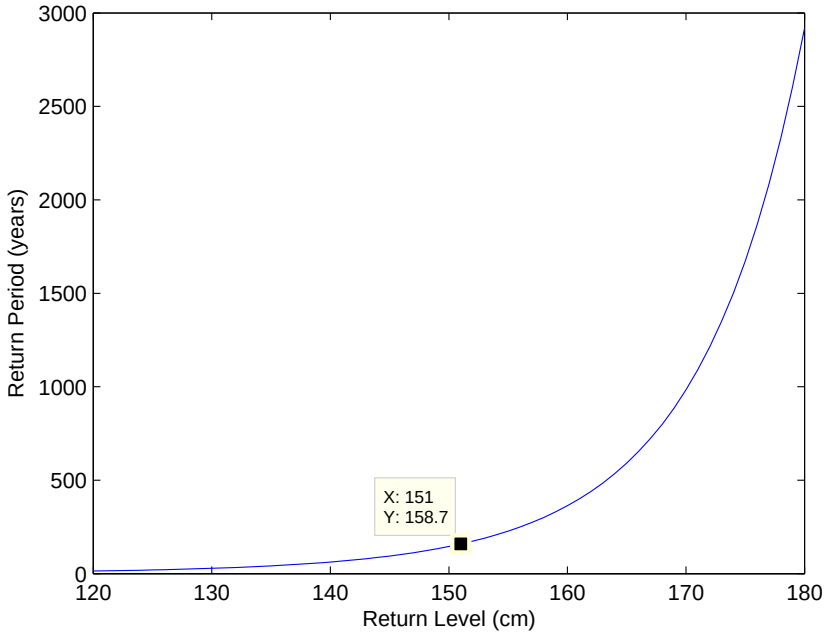
Alla kuvaan 2.11 on piirretty 1 – 10 000 vuoden tapahtumia vastaavat tasot 95 %:n luottamusväleineen. x-akselin asteikko on logaritminen pitkien aikavälien hahmottamisen helpottamiseksi. Kuvaan on piirretty lisäksi alkuperäiset havainnot, jolloin toistumistasokuvaaja voidaan käyttää myös diagnostisena työkaluna mallin hyvyyttä arvioitaessa. Estimoituun malli perustuva kuvaaja näyttää vastaavan varsin hyvin havaittuja toistumistasoja.

Tässä yhteydessä on hyvä muistaa, että tulkitessa toistumistasokuvaajaa pitkillä toistumisperiodeilla (eli pienillä p), vahvojen johtopäätösten kanssa täytyy olla hyvin varovainen. Tämä pätee kaikkeen datan ulkopuolelle ekstrapolointiin. Parametriestimaatit ja niiden luottamusvälit perustuvat oletukseen tai nollahypoteesiin, että malli on oikea. Vaikka mallidiagnostiikka ei anna aihetta hylätä nollahypoteesia – eli malli on tässä mielessä kohtuullinen tai riittävä (adequate) – ei se myöskään todista mallia oikeaksi. Malliin liittyvää epävarmuutta ja saatuja luottamusvälejä tulisikin siten pitää oikestaan alarajoina, tai malliriski (model risk) tulisi pyrkiä ottamaan eksplisiittisesti huomioon esimerkiksi Bayesilaista viitekehystä hyödyntäen.

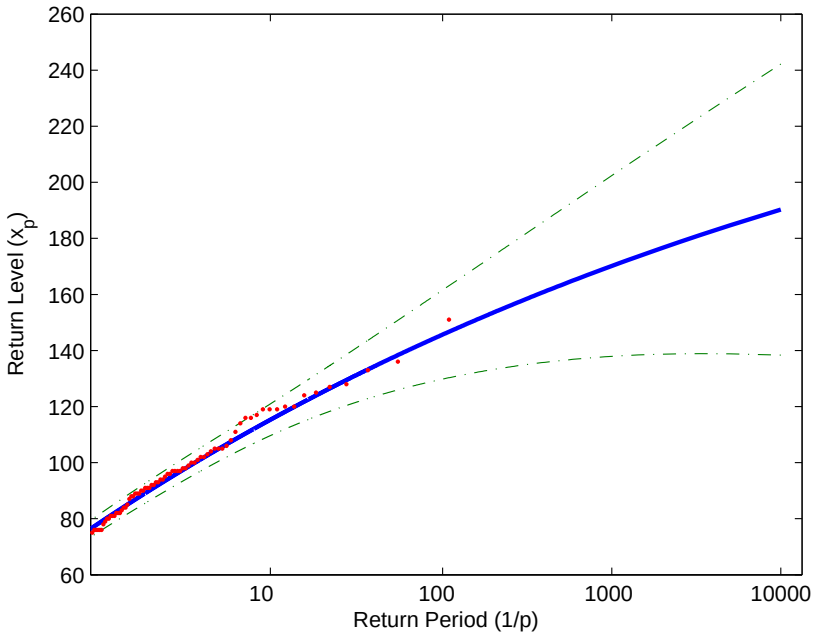
[10] Vrt. osiot 2.5 ja 2.6.



Kuva 2.9: Toistumistasokuvaaja vedenkorkeudelle GEV-mallissa.



Kuva 2.10: Toistumisperiodikuvaaja vedenkorkeudelle GEV-mallissa.



Kuva 2.11: Vedenkorkeuden toistumistasokuvaaja GEV-mallissa luottamusväleineen.

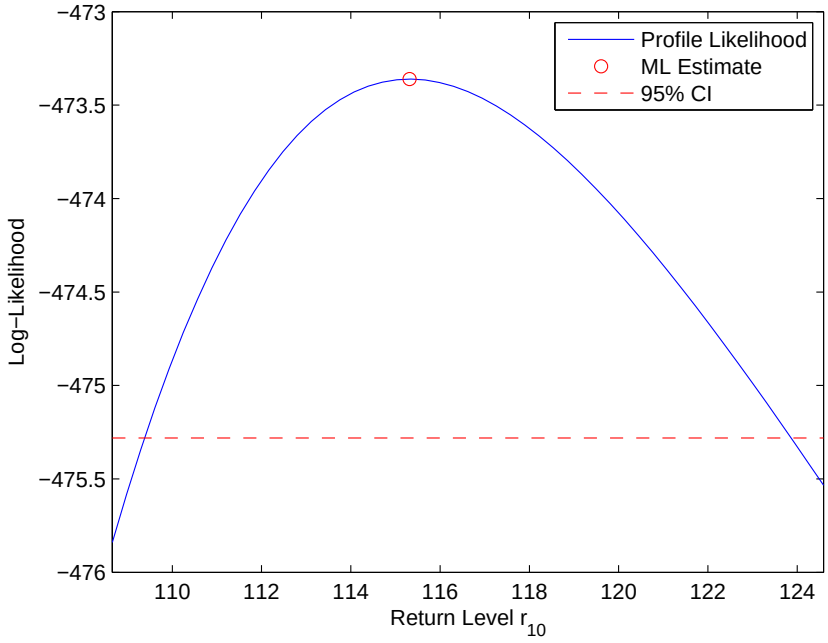
Suurimman uskottavuuden estimaattorin jakauman normaalisuus (joka on siis asyptoottinen tulos otoskoon kasvaessa äärettömiin) ei myöskään välttämättä päde kovin hyvin pienillä otoksilla. Parempaan tarkkuuteen luottamusväleissä päästäänkin yleensä käyttämällä profiiliuskottavuuteen perustuvaa menetelmää (ks. osio 2.2.2.1 ja liite B). Kuvissa 2.12, 2.13, 2.14 ja 2.15 on esitetty profiiliuskottavuusfunktiot 10, 100, 1 000 ja 10 000 vuoden toistumistasoille sekä vastaavat SU-piste-estimaatit. 95 %:n luottamusvälin vasen ja oikea piste saadaan profiiliuskottavuusfunktion ja kriittistä χ^2-arvoa vastaavan horisontaalisen katkoviivan leikkauspisteistä.

Taulukkoon 2.5 alla on koottu tarkasteltuja toistumisperiodeja vastaavat luottamusvälit SU-estimaattorin asymptoottiseen normaalisuuteen ja delta-menetelmään sekä profiiliuskottavuuteen perustuen.

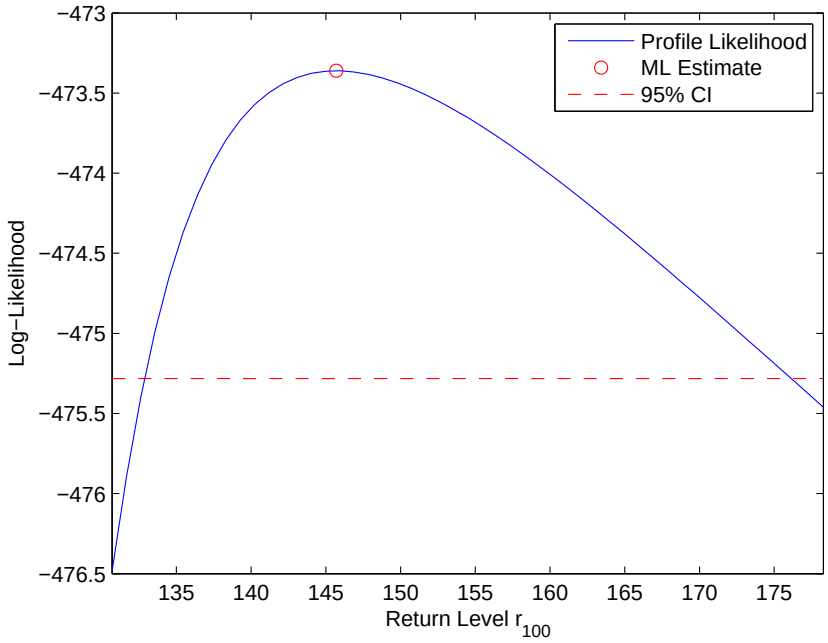
Taulukko 2.5: Delta-menetelmään ja profiliuskottavuuteen perustuvat 95 %:n luottamusvälit merenpinnan korkeuden toistumistasoille GEV-mallissa.

	Delta-menetelmä		**Profiiliuskottavuus**	
Toistumisperiodi (v)	SUE (cm)	95 % CI (cm)	SUE (cm)	95 % CI (cm)
10	115	[109, 121]	115	[109, 124]
100	146	[129, 162]	146	[132, 177]
1 000	170	[137, 203]	170	[146, 241]
10 000	190	[138, 243]	190	[154, 319]

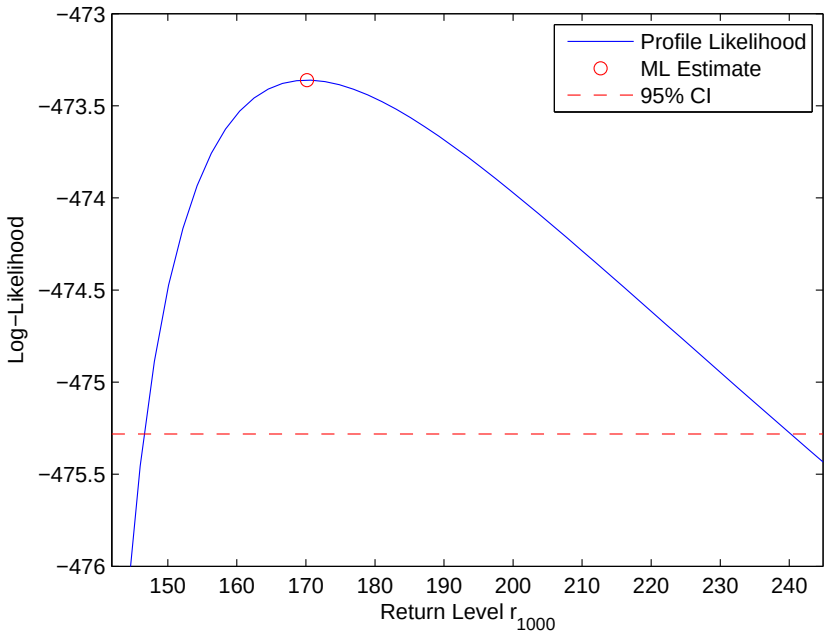
Eri menetelmien tuottamia luottamusvälejä tarkastellessa havaitaan, että ly-



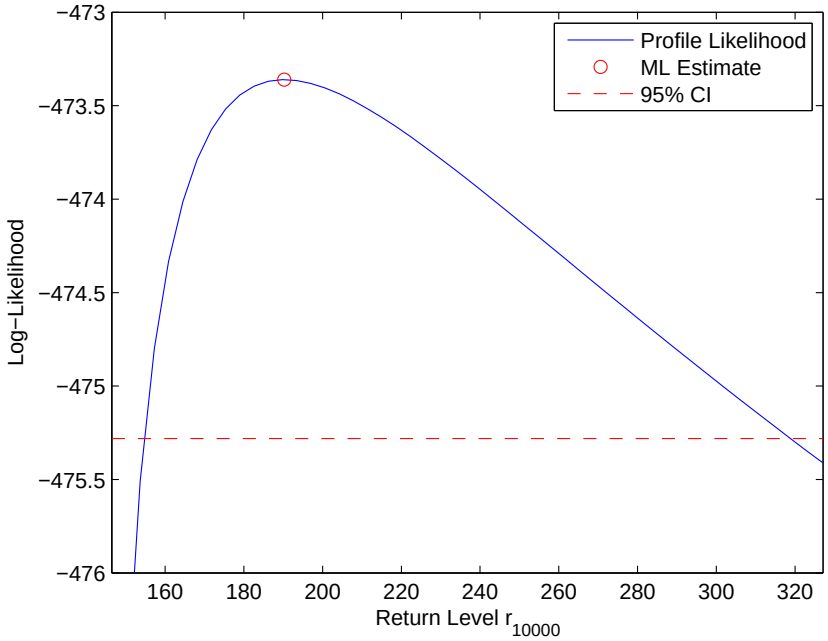
Kuva 2.12: Profiiliuskottavuus merenpinnan korkeuden 10-vuoden toistumistasolle GEV-mallissa.



Kuva 2.13: Profiiliuskottavuus merenpinnan korkeuden 100-vuoden toistumistasolle GEV-mallissa.



Kuva 2.14: Profiiliuskottavuus merenpinnan korkeuden 1 000-vuoden toistumistasolle GEV-mallissa.



Kuva 2.15: Profiiliuskottavuus merenpinnan korkeuden 10 000-vuoden toistumistasolle GEV-mallissa.

hyellä 10-vuoden periodilla nämä ovat samankaltaisia. Tarkastelujakson pidentyessä eli pidemmälle jakauman häntään edetessä ero alkaa kuitenkin kasvaa suureksi. Siinä missä kovarianssimatriisiin perustuvat delta-menetelmän antamat luottamusvälit ovat välttämättä symmetrisiä, profiiliuskottavuusfunktion asymmetrisyys kasvaa voimakkaasti toistumisperiodin kasvaessa, kuten kuvista edellä nähdään. Tämä on luonnollista, sillä havaintoaineisto tarjoaa sitä vähemmän informaatiota vedenkorkeusprosessin jakaumasta, mitä suurempia tasoja (eli pidempiä periodeja) tarkastellaan.

2.2.4.1 Kuukausikohtaiset vuosimaksimit

Tarkastellaan vielä havainnollistuksen vuoksi lyhyesti kuukausikohtaista estimointia perustuen kuukausittaisiin maksimihavaintoihin vuosilta 1904–2011. Kuten vuosikohtaisista maksimeista edellä, kustakin kuukausikohtaisesta vuosittaisesta maksimista on myös 108 havaintoa. Taulukossa 2.6 on esitetty dataan sovitettujen mallien parametriestimaatit.

Taulukko 2.6: Paremetrien SU-estimaatit luottamusväleineen kuukausikohtaisissa GEV-malleissa.

Parametri	$\boldsymbol{\xi}$		$\boldsymbol{\mu}$	
Kuukausi	SUE	95% CI	SUE	95% CI
Tammi	-0.0499	$[-0.186, 0.0863]$	44.8	$[39.3, 50.2]$
Helmi	-0.211	$[-0.345, -0.0761]$	29.1	$[22.9, 35.2]$
Maalis	-0.0173	$[-0.146, 0.112]$	17.4	$[12.3, 22.5]$
Huhti	-0.123	$[-0.234, -0.0111]$	14.1	$[10.6, 17.7]$
Touko	-0.219	$[-0.324, -0.114]$	12.0	$[9.0, 14.9]$
Kesä	-0.163	$[-0.309, -0.0163]$	18.9	$[16.4, 21.5]$
Heinä	-0.144	$[-0.238, -0.0491]$	27.1	$[24.6, 29.6]$
Elo	-0.270	$[-0.385, -0.155]$	32.6	$[29.3, 35.9]$
Syys	-0.321	$[-0.425, -0.216]$	38.5	$[34.1, 42.9]$
Loka	-0.339	$[-0.461, -0.217]$	43.3	$[38.1, 48.6]$
Marras	-0.167	$[-0.284, -0.0495]$	46.9	$[42.0, 51.8]$
Joulu	-0.323	$[-0.445, -0.201]$	51.6	$[45.7, 57.6]$
Parametri	$\boldsymbol{\sigma}$		$\boldsymbol{x_F}$	
Kuukausi	SUE	95% CI	SUE	95% CI
Tammi	25.6	$[21.9, 29.8]$	557	$[-596, 1710]$
Helmi	28.9	$[24.8, 33.7]$	166	$[102, 230]$
Maalis	24.1	$[20.7, 28]$	1410	$[-7220, 10^4]$
Huhti	17.0	$[14.6, 19.7]$	152	$[53, 251]$
Touko	14.0	$[12.1, 16.2]$	76	$[53, 99]$
Kesä	11.8	$[10.1, 13.9]$	92	$[41, 142]$
Heinä	12.0	$[10.4, 13.9]$	111	$[67, 153]$
Elo	15.8	$[13.7, 18.4]$	91	$[74, 108]$
Syys	21.4	$[18.5, 24.7]$	105	$[91, 119]$
Loka	25.1	$[21.6, 29.2]$	117	$[101, 134]$
Marras	23.3	$[20.1, 26.9]$	186	$[111, 261]$
Joulu	28.4	$[24.5, 32.9]$	140	$[118, 161]$

Kuten odotettua (vrt. esim. kuva 2.4), parametriestimaattien arvoissa on pal-

jon eroja eri kuukausien välillä. Toisaalta ne ovat luonteeltaan samankaltaisia. Kaikkien kuukausikohtaisten mallien muotoparametrin ξ estimaatti on negatiivinen, mikä viittaa Weibull-tyypin ääriarvojakaumaan äärellisellä oikealla päätepisteellä. Tammikuun ja maaliskuun kohdalla ξ:n 95 %:n luottamusväli tosin ulottuu positiiviseksi, sisältäen nollan. Aina kun $\xi > 0$, ja yleensä kun $\xi = 0$, on jakauman oikea päätepiste $x_F = \infty$. Tällaisissa tapauksissa oikean päätepisteen x_F tarkastelu ei siis ole mielekästä. Lähellä nollaa olevat muotoparametrin arvot heijastuvatkin taulukossa 2.6 tammikuun ja erityisesti maaliskuun x_F:n SU-estimaatteihin ja luottamusväleihin. Luottamusvälit perustuvat delta-menetelmään, eli niitä vaivaavat aiemmin mainitut ongelmat (tässä luottamusvälin symmetrisyydestä johtuen alaraja menee negatiiviseksi).

Kun ollaan kiinnostuneita vain vedenkorkeuden maksimitasoista ja niiden todennäköisyyksistä – eikä erityisesti niiden sattumisajankohdista – kuten kiinteiden tulvavallien mitoituksen tai omaisuusvahingon todennäköisyyksien estimoinnin tapauksessa, ei kuukausikohtaista tarkastelua tarvita. Sen sijaan kuukausi- tai kausikohtainen tarkastelu voi auttaa ilmiön fysikaalisten ominaisuuksien ymmärtämisessä, etenkin, jos tilastolliseen malliin otetaan mukaan selittäviä muuttujia (ks. osiot 2.5 ja 2.6). Tällöinkään ei yleensä ole perusteltua tarkastella kuukausia erikseen, vaan malliin tulisi rakentaa sisään esimerkiksi aikariippuvat parametrit ja soveltaa sitä suoraan kaikki kuukaudet sisältävään datajoukkoon; ks. edelleen osio 2.5.

2.3 Ylitemenetelmä

Yksi blokkimaksimimenetelmän ongelmista on se, että menetelmä ei välttämättä hyödynnä saatavilla olevaa – ja yleensä jo valmiiksi niukkaa – dataa tehokkaasti. Vain kunkin havaintoblokin suurin havainto huomioidaan mallinnuksessa, vaikka tietyssä blokissa saattaa olla useita havaintoja, jotka ovat suurempia kuin jonkin toisen blokin maksimi. Mikäli data on saatavilla muodossa, joka sisältää muutakin dataa ääri-ilmiöistä kuin vain kunkin havaintojakson maksimin, päästään parempaan lopputulokseen yleensä tarkastelemalla kaikkia määrätyn korkean tason ylittäviä havaintoja. Tämä tarkoittaa tason u ylittävien havaintojen jakauman (ylitejakauman) F_u mallintamista, ja lähestymistapaa kutsutaan tässä esityksessä ylitemenetelmäksi (method of threshold excesses). Esimerkiksi siinä missä blokkimaksimimenetelmä vedenkorkeusdataan sovellettuna tavanomaisesti tarkastelee päivähavainnoista poimittuja vuosittaisia maksimeja, ylitemenetelmässä tarkasteluun otetaan kaikki määrätyn tason ylittävät päivähavainnot.

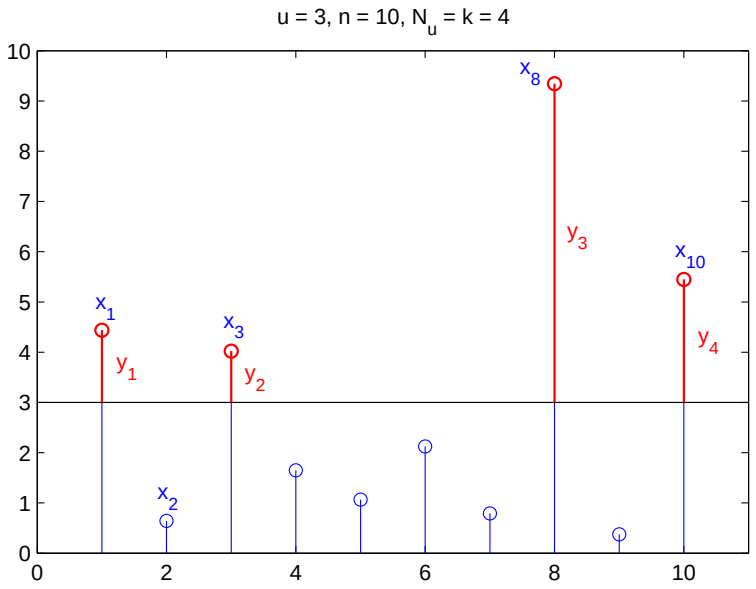
Menetelmän tilastollista implementointia ajatellen lause 1.47 antaa välittömästi seuraavan approksimaation ylitejakaumalle F_u, suurilla u:

$$F_u(x) = \mathbb{P}(X - u \leq x | X > u) \approx G_{\xi,\beta(u)}(x), \quad x > 0.$$

Oletetaan, että tarkasteltava otos koostuu satunnaismuuttujajonosta $X_1, \ldots, X_n$ yhteisenä kertymäfunktiona F. Näistä määrä N_u ylittää tason u; N_u on luonnollisesti satunnaismuuttuja. Uudelleennimetään tason u ylittävät havainnot $\tilde{X}_1, \ldots, \tilde{X}_{N_u}$, ja merkitään ylitteiden suuruuksia

$$Y_j = \tilde{X}_j - u, \quad j = 1, \ldots, N_u;$$

ks. kuva 2.16. Jos tarkastellaan yhtä aikaa useita eri tasoja u ja on olemassa sekaannuksen vaara, merkitään kynnystaso eksplisiittisesti näkyviin kirjoittamalla $\tilde{X}^{(u)}$, $Y_j^{(u)}$. Ylitemenetelmä koostuu pääpiirteissään kynnyksen tai tason u valinnasta ja yleistetyn Pareto-jakauman $G_{\xi,\beta}$ sovittamisesta tason ylitteisiin $(Y_j)_{j=1}^{N_u} = (Y_j^{(u)})_{j=1}^{N_u}$.



Kuva 2.16: Ylitemenetelmän datan havainnollistus: havaittu otos $\boldsymbol{X} = \boldsymbol{x}$ ja tason $u = 3$ ylitteet $\boldsymbol{Y} = \boldsymbol{y}$

2.3.1 Kynnystason valinta

Kuten blokkimenetelmässä blokin koon valinnan yhteydessä, myös ylitemenetelmän kynnystasoa valittaessa joudutaan tasapainottelemaan parametriestimaattien varianssin ja harhan välillä: liian suuri kynnyksen u arvo johtaa siihen, että estimointia varten jää vain muutama havainto ja siten estimaattorien varianssi on liian suuri, kun taas liian pienen kynnyksen valinta johtaa siihen, että ylitejakauman GPD-approksimaatio on huono, ja estimaattoreista tulee harhaisia. Ylitemenetelmän kohdalla valintaongelma on itse asiassa käytännössä vaikeampi kuin blokkimaksimimenetelmän kohdalla, koska jälkimmäisessä voidaan usein osoittaa jossain mielessä luonnollinen datablokin koko, kuten vuosi monien luonnonilmiöiden kohdalla. Ylitemenetelmän osalta tällaista luonnollista kynnysarvoa harvemmin on olemassa.

Teoreettisesti on mahdollista valita kynnys u asymptoottisesti optimaalisesti tarkastelemalla eksplisiittisesti kompromissia harhan ja varianssin kesken [2, s. 355], mutta käytännössä yksiselitteistä ratkaisua ongelmaan ei ole esitetty. Käytännössä ongelmaa lähestytäänkin sovelluksien yhteydessä yleensä pragmaattisesti pyrkimällä valitsemaan mahdollisimman matala kynnys siten, että

ylitejakauman aproksimointi yleistetyllä Pareto-jakaumalla vaikuttaa vielä perustellulta. Valinnan tueksi on esitetty kaksi GP-jakauman ominaisuuksiin perustuvaa menetelmää, jotka esitetään seuraavaksi: osiossa 1.5 tavattu ylitteen odotusarvofunktio, sekä parametriestimaattien stabiilisuuden arviointi sovittamalla malli useita eri kynnysarvoja käyttäen.

2.3.1.1 Ylitteen odotusarvofunktio

Oletetaan, että $\xi < 1$, eli yleistetyn Pareto-jakauman odotusarvo on olemassa (mikäli näin ei ole – eli hyvin paksuhäntäisten jakaumien tapauksessa – ei tämän osion menetelmä suoraan sovellu). Määritelmästä 1.46 muistetaan ylitteen odotusarvofunktio,

$$e(u) = \mathbb{E}(X - u | X > u).$$

Oletetaan, että yleistetty Pareto-jakauma pätee jonon $X_1, \ldots, X_n$ tuottamille tason u_0 ylitteille, jolloin

$$e(u_0) = \frac{\beta_{u_0}}{1 - \xi},$$

missä on merkitty tasoa u_0 vastaavaa skaalaparametria $\beta = \beta_{u_0}$. Mutta jos GPD on pätevä malli tason u_0 ylitteille, se on sitä myös kaikille korkeamman tason $u > u_0$ ylitteille. Osion 1.5 mukaisesti saadaan tällöin, kun $u > u_0$,

$$e(u) = \frac{\beta_u}{1 - \xi} = \frac{\beta_{u_0} + \xi u}{1 - \xi} = \frac{\beta_{u_0}}{1 - \xi} + \frac{\xi}{1 - \xi} u.$$

Ylitteen odotusarvofunktio on siis u:n lineaarinen funktio, kun $u > u_0$. Kynnyksen valinta voidaan perustaa tähän GP-jakauman ominaisuuteen seuraavaksi kuvatulla tavalla.

Ylitteen odotusarvofunktio $e(u) = \mathbb{E}(X - u | X > u)$ on nimensä mukaisesti tason u ylitteiden ehdollinen odotusarvo, ja sen empiirinen vastine on tason u ylittävien realisaatioiden otoskeskiarvo,

$$\hat{e}(u) = \frac{1}{\#\{i : X_i > u\}} \sum_{\{i : X_i > u\}} (X_i - u) = \frac{1}{N_u} \sum_{i=1}^{N_u} Y_i^{(u)}.$$

Estimaatin $\hat{e}(u)$ tulisi kasvaa likimain lineaarisesti (otoshajonnan huomioimisen jälkeen) u:n suhteen niillä tasoilla u, joilla GP-jakauma tarjoaa sopivan mallin havainnoille. Kynnyksen valintaongelmaa voidaan siis lähestyä graafisesti piirtämällä pisteet

$$\{(u, \hat{e}(u)) : u < X_{max}\},$$

missä $X_{max} = \max(X_i, \ldots, X_n)$ ja etsimällä datasta aluetta, josta alkaen kuvaaja on likimain lineaarinen. Kuvaajaa kutsutaan ylitteen otoskeskiarvokuvaajaksi (sample mean excess plot) tai, erityisesti luotettavuustekniikan piirissä, jäljellä olevan keskieliniän kuvaajaksi (mean residual life plot). Otoskeskiarvolle $\hat{e}(u)$ voidaan muodostaa luottamusvälit perustuen otoskeskiarvojen likimääräiseen normaalijakautuneisuuteen.

2.3.1.2 Parametrien stabiilisuus

Toinen menetelmä kynnystason u valitsemiseksi perustuu parametriestimaattien stabiilisuuden tarkasteluun, kun mallin sovitus tehdään lukuisilla eri u:n arvoilla. Lähestymistapa pohjaa olennaisesti samaan argumenttiin kuin ylitteen odotusarvofunktion käyttökin. Mikäli yleistetty Pareto-jakauma on pätevä malli kynnyksen u_0 ylitteille, niin myös tätä tasoa korkeamman tason u ylitteet noudattavat osion 1.5 ja edellä sanotun mukaisesti GP-jakaumaa samalla muotoparametrilla ξ, mutta skaalaparametrilla

$$\beta_u := \beta(u - u_0) = \beta_{u_0} + \xi(u - u_0),$$

missä β_{u_0} on tasoa u_0 vastaava skaalaparametri. Parametrin ξ pitäisi siis pysyä vakiona (otosvarianssin huomioimisen jälkeen käytännössä stabiilina) tason u_0 jälkeen, mikäli GPD on sopiva malli tason u_0 ylitteille. Vastaavasti skaalaparametrin β_u tulisi muuttua likimain lineaarisesti tason u mukana, paitsi jos $\xi = 0$. Tästä hankaluudesta päästään eroon ottamalla käyttöön vaihtoehtoinen parametrisaatio

$$\beta^* := \beta_u - \xi u,$$

joka on siis vakio u:n suhteen ($\beta^* = \beta_{u_0} - \xi u_0$).

Kynnyksen valintaa varten piirretään siis estimaatit $\hat{\xi}$ ja $\hat{\beta}^*$ tasoa u vasten, ja valitaan kynnykseksi u_0 pienin u, josta lähtien parametrit pysyvät likimain vakioina. Käytännössä kuvaan on välttämätöntä piirtää mukaan myös luottamusvälit, jotta parametrien vakioisuudesta voidaan perustellusti sanoa jotain. Luottamusvälit $\hat{\xi}$:lle (ja $\hat{\beta}_u$:lle) saadaan tutusti SU-estimaattorien asymptoottiseen normaalisuuteen perustuen parametriestimaattien kovarianssimatriisista $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}$. Muunnetun parametrin $\hat{\beta}^*$ kohdalla täytyy käyttää delta-menetelmää, jolloin likimääräiseksi varianssiksi saadaan

$$\mathrm{Var}(\hat{\beta}^*) = \nabla\beta^{*T}\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}\nabla\beta^*,$$

missä $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}$ on parametriestimaattien $\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\beta}_u)$ (asymptoottinen) kovarianssimatriisi, ja gradientti on

$$\nabla\beta^{*T} = \begin{pmatrix} \dfrac{\partial\beta^*}{\partial\xi} \\ \dfrac{\partial\beta^*}{\partial\beta_u} \end{pmatrix} = \begin{pmatrix} -u \\ 1 \end{pmatrix}.$$

2.3.2 Suurimman uskottavuuden menetelmä GP-jakaumalle

Kun kynnystaso u on valittu, voidaan GP-jakauma sovittaa dataan eli jakauman parametrit estimoida suurimman uskottavuuden menetelmällä. GPD:n tiheysfunktio parametreilla $\xi \in \mathbb{R}$ ja $\beta > 0$ on

$$g_{\boldsymbol{\theta}}(x) = \begin{cases} \dfrac{1}{\beta}\left(1 + \xi\dfrac{x}{\beta}\right)^{-\frac{1}{\xi}-1}, & \xi \neq 0, \\ e^{-x/\beta}, & \xi = 0, \end{cases} \quad , \quad x \in D(\boldsymbol{\theta}),$$

missä määrittelyalue on

$$D(\boldsymbol{\theta}) = \begin{cases} [0, \infty), & \text{kun} \quad \xi \geq 0, \\ [0, -\beta/\xi), & \text{kun} \quad \xi < 0, \end{cases}$$

ja parametrien muodostama vektori $\boldsymbol{\theta} = (\xi, \beta)$.

Merkitään tason u ylittäviä havaintoja $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_{N_u})$ ja vastaavia ylitteitä $\boldsymbol{Y} = (Y_1, \ldots, Y_{N_u})$, missä siis $Y_j = \tilde{X}_j - u$. Oletetaan, että Y_i ovat iid, $Y_i \sim G_{\xi,\beta}\ \forall\ i = 1, \ldots, N_u$. Olkoon havaittu $N_u = k$ tason u ylitystä. Havaittuun otokseen $\boldsymbol{Y} = \boldsymbol{y} = (y_1, \ldots, y_k)$ perustuva logaritminen uskottavuusfunktio on tällöin (kun $\xi \neq 0$)

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{k} \ln g_{\boldsymbol{\theta}}(y_i) = -k \ln \beta - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k} \ln \left(1 + \xi \frac{y_i}{\beta}\right), \tag{2.11}$$

kun $y_i \in D(\xi, \beta)$ eli $1 + \xi y_i/\beta \geq 0$ kaikilla $i = 1, \ldots, k$. Mikäli määrittelyalueen asettama rajoitusehto ei täyty jollain i, on $l(\boldsymbol{\theta}; \boldsymbol{y}) = -\infty$.

Tapauksessa $\xi = 0$ saadaan log-uskottavuusfunktioksi

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = -k \ln \beta - \frac{1}{\beta} \sum_{i=1}^{k} y_i, \tag{2.12}$$

missä nyt siis $\boldsymbol{\theta} = \theta = \beta$, vastaten eksponenttijakaumaa parametrilla $1/\beta$. Sekä tapauksessa $\xi = 0$ että $\xi \neq 0$ täytyy luonnollisesti olla $\beta > 0$. Tämä on jälleen helpointa varmistaa käyttämällä numeerisen laskennan toteutuksessa β:n sijasta muunnosta $\tilde{\beta} := \ln \beta$, joka on aina positiivinen ja voidaan lopuksi muuntaa takaisin eksponenttiin korottamalla, $\beta = \exp(\tilde{\beta})$.

Parametriestimaattien keskivirheet ja sitä kautta luottamusvälit saadaan asymptoottisesta kovarianssimatriisista tutusti aiemmin esitetyllä tavalla. Profiiliuskottavuuteen perustuvat luottamusvälit saadaan myös kuten aiemmin, maksimoimalla (logaritminen) profiiliuskottavuusfunktio muiden (toisen) parametrien suhteen.

2.3.3 Toistumisperiodi ja toistumistaso

Tarkastellaan seuraavaksi toistumistasoja ja –periodeja ylitemenetelmässä. Toistumistason määräämiseksi täytyy huomioida ylitteiden jakauman lisäksi todennäköisyys, jolla ylitteet tapahtuvat. Muistetaan, että ylitejakauma on

$$F_u(y) = \mathbb{P}(X - u \leq y | X > u) = \mathbb{P}(Y \leq y | X > u), \quad y \leq 0,$$

missä $Y = X - u$. Tapahtuman $\{Y > y\} = \{X > u + y\}$ ehdollistamaton todennäköisyys voidaan kirjoittaa

$$\mathbb{P}(Y > y) = \mathbb{P}(X > u)\mathbb{P}(Y > y | X > u),$$

tai lyhyemmin $\bar{F}(u + y) = \bar{F}(u)\bar{F}_u(y)$. Nyt tason u ylitejakauma on yleistetty Pareto-jakauma, $\bar{F}_u(y) = \bar{G}_{\boldsymbol{\theta}}(y)$, ja tapahtuman $\{Y > y\}$ todennäköisyydeksi tulee

$$\mathbb{P}(Y > y) = \zeta_u \bar{G}_{\boldsymbol{\theta}}(y),$$

missä tason u ylittämisen todennäköisyyttä on merkitty lyhyesti $\zeta_u := \mathbb{P}(X > u)$.

Keskimäärin kerran m:ssä havainnossa ylitettävä taso x_m $(> u)$ saadaan nyt yhtälöstä $\mathbb{P}(Y > x_m - u) = 1/m$, eli, kun $\xi \neq 0$,

$$\zeta_u \left(1 + \xi \frac{x_m - u}{\beta}\right)^{-1/\xi} = \frac{1}{m}.$$

Tämä kääntämällä saadaan ratkaistua

$$x_m = u + \frac{\beta}{\xi}\left((m\zeta_u)^{\xi} - 1\right),$$

kun $x_m > u$. Tapauksessa $\xi = 0$ saadaan vastaavasti tulokseksi

$$x_m = u + \beta \ln(m\zeta_u),$$

kun $x_m > u$.

Edellä saatu x_m on m. havainnon toistumistaso, eli taso, joka ylitetään odotusarvoisesti kerran m:ssä havainnossa.[11] Tuloksien tulkintaa ja esittämistä varten on usein mukavampi tarkastella toistumistasoja jotka on määritelty vuositasolla, siten että N-vuoden toistumistaso on taso, joka keskimäärin ylitetään kerran N:ssä vuodessa. Oletetaan siis, että vuodessa on n_y havaintoa, jolloin N-vuoden toistumistaso vastaa m-havainnon toistumistasoa arvolla $m = n_y N$. Tällöin N-vuoden toistumistaso on

$$x_N = u + \frac{\beta}{\xi}\left((n_y N\zeta_u)^{\xi} - 1\right), \tag{2.13}$$

kun $\xi \neq 0$, ja

$$x_m = u + \beta \ln(n_y N\zeta_u), \tag{2.14}$$

kun $\xi = 0$.

Toistumistasoilla saadaan estimaatti, kun estimoidut parametriarvot sijoitetaan yo. yhtälöihin. Todennäköisyyden $\zeta_u = \mathbb{P}(X > u)$ havaittuun otokseen perustuva luonnollinen estimaatti on tason u ylittävien havaintojen (k kpl) osuus kaikista havainnoista (n kpl),

$$\hat{\zeta}_u = \frac{k}{n}.$$

Toisaalta ylitystapahtuman indikaattori $\mathbb{1}_{\{X_i > u\}}$ on Bernoulli-jakautunut satunnaismuuttuja onnistumistodennäköisyydellä $p = \mathbb{P}(X > u) = \zeta_u$, jolloin ylitysten lukumäärä $N_u = \sum_{i=1}^{n} \mathbb{1}_{\{X_i > u\}}$ on binomijakautunut parametrillä (n, ζ_u). Tästä seuraa, että $\hat{\zeta}_u$ on myös todennäköisyyden ζ_u suurimman uskottavuuden estimaattori.

Luottamusvälit. Keskivirheet ja niihin perustuvat luottamusvälit toistumistasolle saadaan johdettua normaaliin tapaan delta-menetelmällä. Nyt kuitenkin

[11] Tai, täsmällisemmin, taso jonka havainto ylittää todennäköisyydellä $1/m$ – iid tapauksessa näillä ei ole eroa.

myös ylitystodennäköisyyden ζ_u estimaattiin $\hat{\zeta}_u$ liittyvä epävarmuus tulee huomioida. ζ_u:n binomijakautuneisuuden nojalla estimaattorin varianssi on

$$\mathrm{Var}\left(\hat{\zeta}_u\right) = \frac{\hat{\zeta}_u(1-\hat{\zeta}_u)}{n},$$

ja parametriestimaattien $\hat{\boldsymbol{\theta}}_\zeta = (\hat{\xi}, \hat{\beta}, \hat{\zeta}_u)$ kovarianssimatriisiksi saadaan

$$\boldsymbol{V}_{\hat{\boldsymbol{\theta}}_\zeta} = \begin{pmatrix} v_{1,1} & v_{1,2} & 0 \\ v_{2,1} & v_{2,2} & 0 \\ 0 & 0 & \hat{\zeta}(1-\hat{\zeta})/n \end{pmatrix},$$

missä $v_{i,j}$ on parametriestimaattien $\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\beta})$ kovarianssimatriisin $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}$ alkio (i,j). Delta-menetelmään perustuen toistumistason x_m (tai x_N, kun m korvataan termillä $n_y N$) estimaatin $\hat{x}_m$ likimääräinen varianssi on

$$\mathrm{Var}(\hat{x}_m) = \nabla x_m^T \boldsymbol{V}_{\hat{\boldsymbol{\theta}}_\zeta} \nabla x_m,$$

missä gradientiksi tulee

$$\nabla x_m = \begin{pmatrix} \frac{\partial x_m}{\partial \xi} \\ \frac{\partial x_m}{\partial \beta} \\ \frac{\partial x_m}{\partial \zeta_u} \end{pmatrix} = \begin{pmatrix} -\frac{\beta}{\xi^2}\left[(m\zeta_u)^\xi - 1\right] + \frac{\beta}{\xi}(m\zeta_u)^\xi \ln(m\zeta_u) \\ \frac{1}{\xi}\left[(m\zeta_u)^\xi - 1\right] \\ \beta m^\xi \zeta_u^{\xi-1} \end{pmatrix},$$

evaluoituna pisteessä $(\hat{\xi}, \hat{\beta}, \hat{\zeta}_u)$.

Delta-menetelmää parempaan tarkkuuteen päästään yleensä jälleen profiiliuskottavuuteen perustuvaa lähestymistapaa käyttäen. Toistumistasoille profiiliuskottavuus voidaan muodostaa uudelleenparametrisoimalla malli; ζ_u:hun liittyvä epävarmuus jätetään tässä yhteydessä tavallisesti huomioimatta tarkastelun helpottamiseksi, sillä se on yleensä pientä muihin parametreihin liittyvään epävarmuuteen verrattuna. [1] Yhtälöistä (2.13) ja (2.14) saadaan ratkaistua esimerkiksi β muiden parametrien funktiona:

$$\beta = \begin{cases} \xi \dfrac{x_m - u}{(m\zeta_u)^\xi - 1}, & \xi \neq 0, \\ \dfrac{x_m - u}{\ln(m\zeta_u)}, & \xi = 0. \end{cases}$$

Sijoittamalla tämä log-uskottavuusfunktioon saadaan siitä yhden parametrin (eli ξ:n) funktio, joka voidaan maksimoida ξ:n suhteen, kiinteällä x_m. Kun menettely toistetaan useille x_m:n arvoille, saadaan muodostettua toistumistason profiiliuskottavuus. Luottamusvälit seuraavat tästä.

2.3.4 Mallidiagnostiikkaa

Kuten GEV-jakauman kohdalla blokkimenetelmässä, tarkastellaan sovitetun mallin hyvyyttä todennäköisyys- ja kvantiilikuvaajien avulla. Oletetaan siis, että on

valittu kynnys u, jonka ylitteet datassa ovat $\boldsymbol{y} = (y_1, \ldots, y_k)$, ja näihin on sovitettu GP-jakauma,

$$\hat{G}_{\boldsymbol{\theta}}(y) = G_{\hat{\boldsymbol{\theta}}}(y) = \begin{cases} 1 - \left(1 + \hat{\xi}\frac{x}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}}, & \hat{\xi} \neq 0, \\ 1 - e^{-x/\hat{\beta}}, & \hat{\xi} = 0, \end{cases}$$

Todennäköisyyskuvaaja koostuu nyt pisteistä

$$\left\{ \left(\frac{i}{k+1}, \hat{G}_{\boldsymbol{\theta}}(y_{i,k}) \right) : i = 1, \ldots, k \right\}, \tag{2.15}$$

missä siis k on tason u havaittujen ylitteiden lukumäärä alkuperäisessä otoksessa $\boldsymbol{x} = (x_1, \ldots, x_n)$, ja $y_{1,k} \leq \cdots \leq y_{k,k}$ on ylitteiden järjestetty otos.

Kvantiilikuvaaja on vastaavasti

$$\left\{ \left(\hat{G}^{-1}_{\boldsymbol{\theta}}(\frac{i}{k+1}), y_{i,k} \right) : i = 1, \ldots, k \right\}, \tag{2.16}$$

missä nyt

$$\hat{G}^{-1}_{\boldsymbol{\theta}}(y) = \begin{cases} \frac{\hat{\beta}}{\hat{\xi}} \left((1-y)^{-\hat{\xi}} - 1 \right), & \hat{\xi} \neq 0, \\ -\hat{\beta} \ln(1-y), & \hat{\xi} = 0, \end{cases}$$

Mikäli GPD on kohtuullinen malli tason u ylitteille, sekä todennäköisyys- että kvantiilikuvaajien tulisi olla lähes lineaarisia eli koostua pisteistä jotka ovat lähellä yksikködiagonaalia.

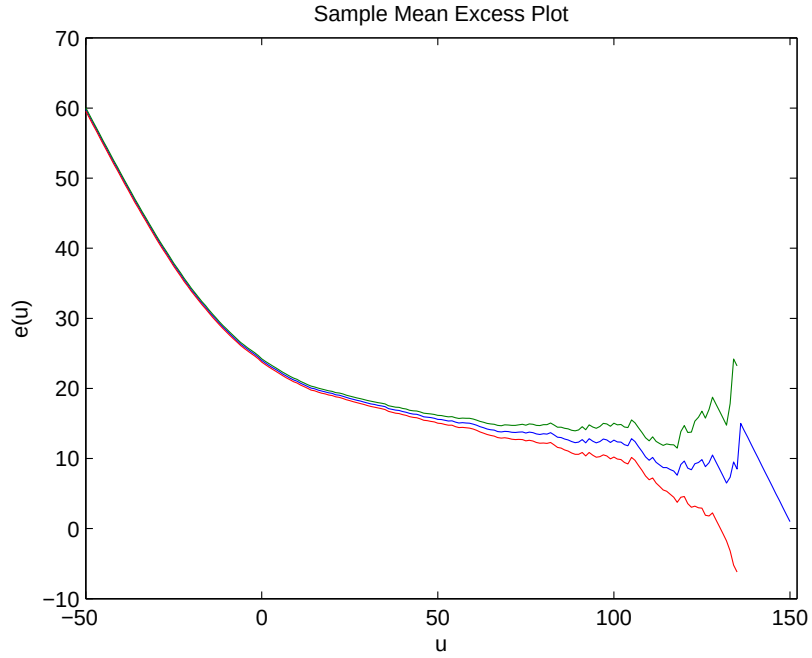
2.3.5 Merenpinnan korkeuden mallintaminen ylitemenetelmällä

Jatketaan Itämeren vedenkorkeuden tarkastelua soveltamalla ylitemenetelmää. Ylitteitä tarkastellessa käytetään nyt koko päivämaksimien havaintoaikasarjaa aikaväliltä 1.1.1904–31.12.2011 (ks. kuva 2.2).

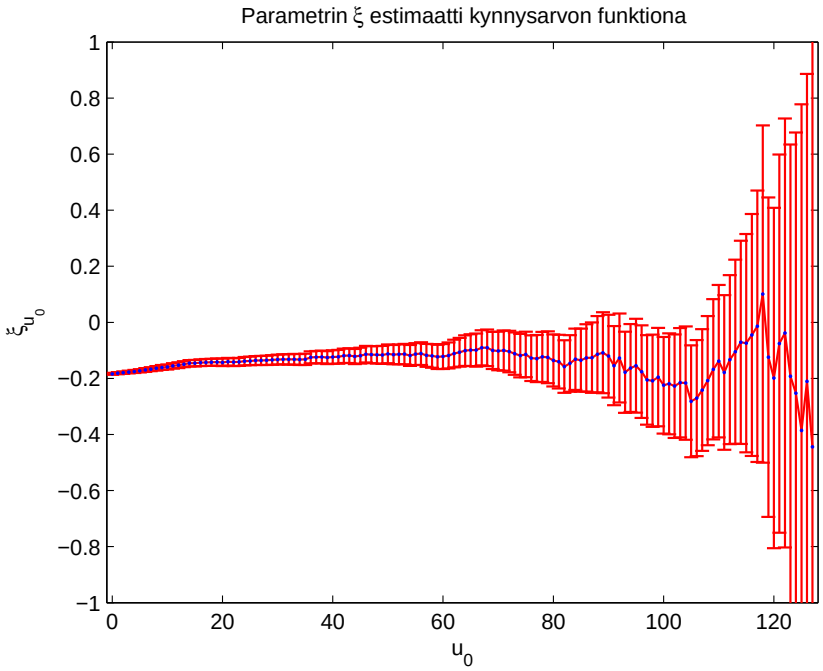
Ennen yleistetyn Pareto-jakauman sovittamista täytyy valita käytettävä kynnystaso u jonka ylittävät havainnot otetaan mukaan tarkasteluun. Kuvassa 2.17 on esitetty ylityksien otoskeskiarvokuvaaja havaintoaineistolle 95 %:n luottamusväleineen.

Nähdään, että kuvaaja kaartuu arviolta kynnysarvoon 94 cm – 96 cm saakka, jonka jälkeen se pysyy tasaisena, kunnes arvon 105 cm kohdalla laskee arvoon 118 cm asti, ja tämän jälkeen heilahtelee ylittävien keskiarvoistettavien havaintojen määrän laskiessa kohti nollaa kynnystason kasvaessa.

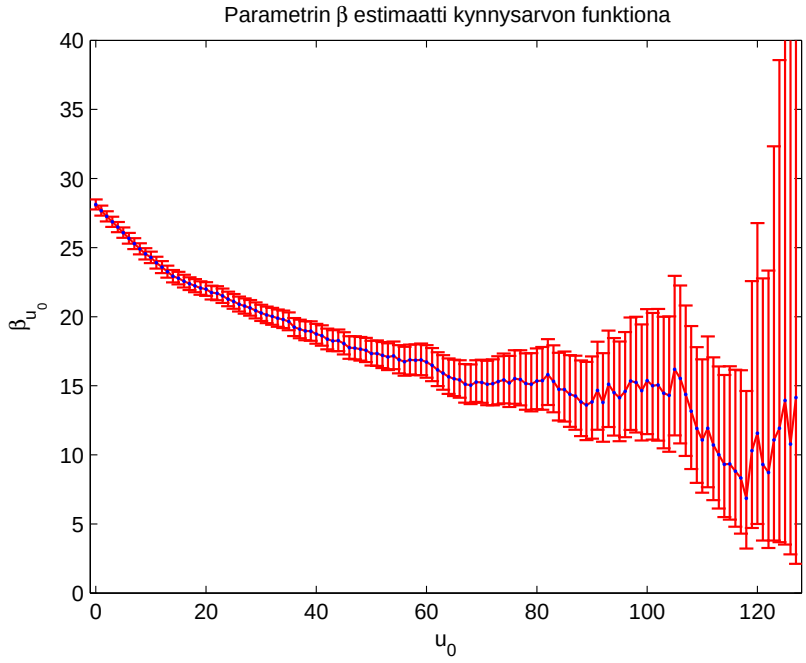
Tarkastellaan lisäksi parametriestimaattien stabiilisuutta kynnystason suhteen estimoimalla GPD-malli toistuvasti kynnysarvoilla $u \in [0, 126]$ (cm); tämän jälkeen ylitteitä on niin vähän, että parametriestimaattien heilunta on epäinformatiivista. Kuvissa 2.18, 2.19 ja 2.20 on piirretty parametriestimaatit kynnystason u funktiona sekä 95 %:n luottamusvälit estimaateille.



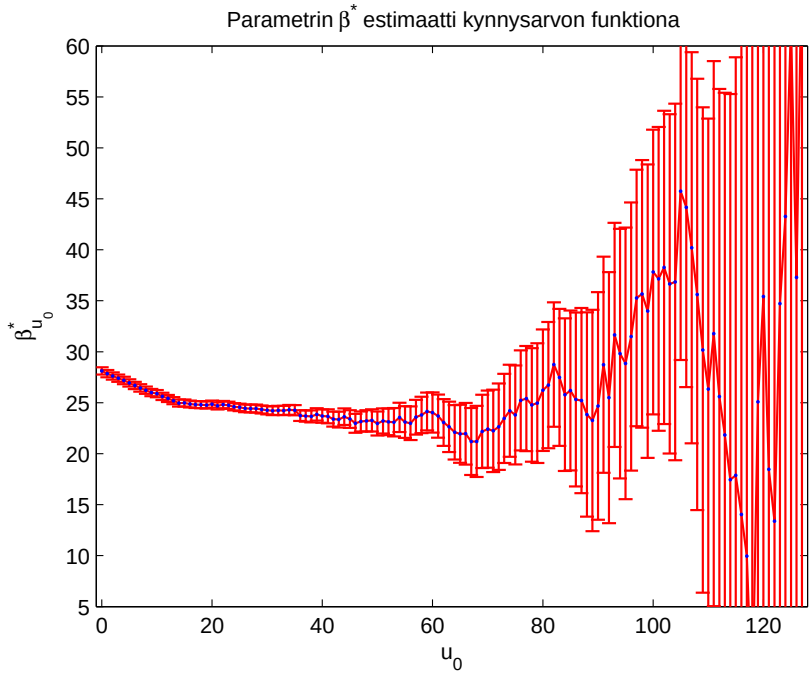
Kuva 2.17: Ylitteen otoskeskiarvokuvaaja vedenkorkeuden päivämaksimidatalle.



Kuva 2.18: Parametrin ξ estimaatti vedenkorkeuden päivämaksimidataan sovitetulle GP-jakaumalle kynnysarvon funktiona.



Kuva 2.19: Parametrin β estimaatti vedenkorkeuden päivämaksimidataan sovitetulle GP-jakaumalle kynnysarvon funktiona.



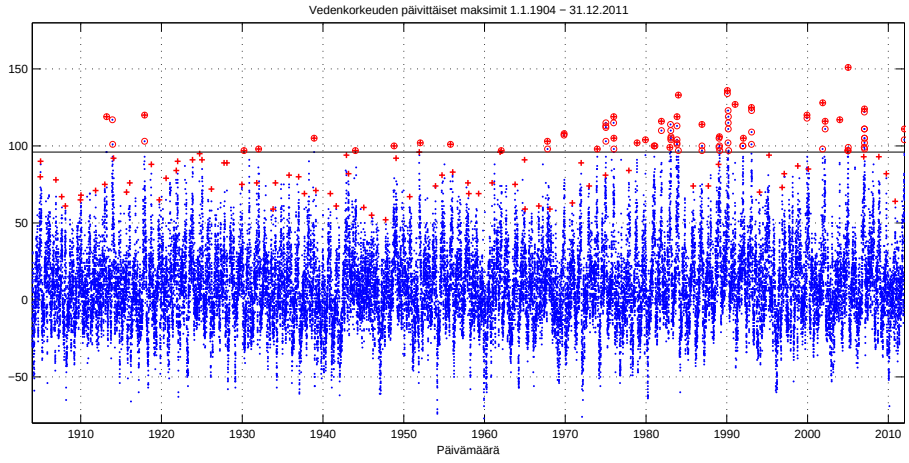
Kuva 2.20: Parametrin β^* estimaatti vedenkorkeuden päivämaksimidataan sovitetulle GP-jakaumalle kynnysarvon funktiona.

Nähdään, että muotoparametrin ξ estimaatti kuvassa 2.18 vaikuttaa itse asiassa kokonaisuutena tarkastellen melko vakiolta koko laajalla tarkasteluvälillä. Alussa on tosin havaittavissa selvää kaartumista. Lopussa nähdään myös ylitteiden määrän vähentyessä tyypillistä heilahtelua. Pelkästään tämän kuvan perusteella kynnystason saattaisi voida luottamusvälit huomioiden asettaa niinkin alhaalle kuin 80 cm tasolle.

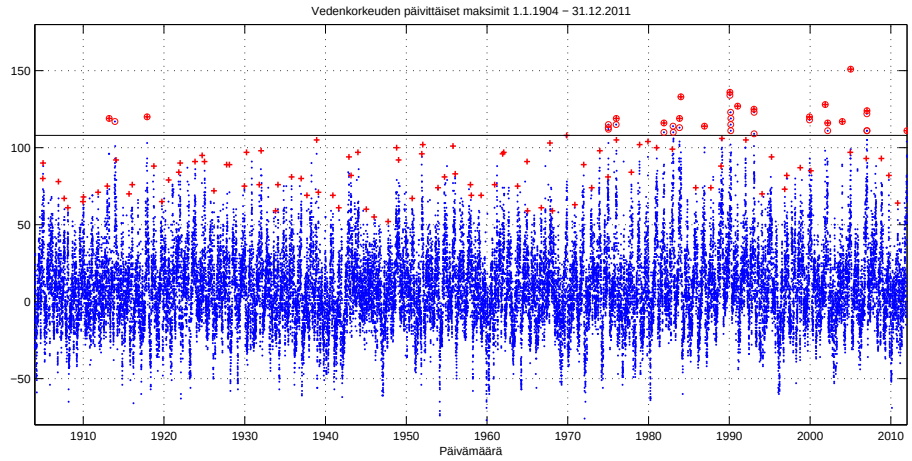
Kuvan 2.19 estimaatin $\hat{\beta}$ tulisi muuttua likimain lineaarisesti u:n funktiona niillä tasoilla, joilla GPD-malli pätee ylitteille, kun $\xi \neq 0$. Kun $\xi = 0$, tulisi estimaatin vastaavasti pysyä likimain vakiona näillä tasoilla. Kuvan tulkintaa vaikeuttaa nyt se, että estimaatin $\hat{\xi}$ kuvaajan mukaan nolla sisältyy ξ:n (95 %:n) luottamusväliin useimmilla korkeilla (eli relevanteilla) kynnystason arvoilla. Kuvaan 2.20 on piirretty muunnetun parametrin $\beta^* = \beta - \xi u$ estimaatti $\hat{\beta}^* = \hat{\beta} - \hat{\xi} u$; tämän tulisi pysyä vakiona ξ:n arvosta riippumatta. Muunnoksen hinta on kuitenkin se, että parametriestimaattiin $\hat{\beta}^*$ tulee epävarmuutta sekä parametrista ξ että parametrista β. Kuvasta nähdään, että estimaatit heilahtelevat voimakkaasti, ja luottamusvälit kasvavat suuriksi. Kuvan 2.20 perusteella 80 cm:n tienoilla oleva kynnysarvo vaikuttaa liian alhaiselta.

Ylitteen otoskeskiarvokuvaajan sekä parametriestimaattien stabiilisuuden tarkastelun perusteella päädytään valitsemaan kynnysarvot $u_1 = 96$ cm ja $u_2 = 108$ cm jatkotarkastelua varten. Nämä on valittu mahdolliselta vaikuttavien arvojen ala- ja yläpäästä, pyrkien kuitenkin varmistamaan, ettei jakauman sovittamista varten jää liian vähän havaintoja. Tason $u = 96$ cm ylittäviä havaintoja on aineistossa $k = 95$ kpl, kun tason $u = 108$ cm ylitteitä on $k = 39$ kpl.

Kuvassa 2.21 on päivittäisten vedenkorkeusmaksimien aikasarjaan merkitty tason $u = 96$ cm (musta poikkiviiva kuvassa) ylittävät havainnot punaisilla palloilla. Kuvaan on myös havainnollisuuden vuoksi merkitty vuosimaksimit punaisilla tähdillä. Nähdään, että iso osa vuosimaksimeista ei ylitä valittua tasoa, kun taas tason ylittävien havaintojen joukossa on paljon sellaisia, jotka eivät ole vuosimaksimeja. Kuvassa 2.22 on vastaava informaatio tason $u = 108$ cm osalta.



Kuva 2.21: Vedenkorkeuden päivämaksimiaikasarja ja tason $u = 96$ cm (musta viiva) ylittävät havainnot.



Kuva 2.22: Vedenkorkeuden päivämaksimiaikasarja ja tason $u = 108$ cm (musta viiva) ylittävät havainnot.

Kuvista nähdään, että ylityksiä ei näytä tapahtuvan tasaisesti. Erityisesti tasoa nostaessa ylitteiden väliin jää vuosikymmenien väli, siten että esimerkiksi korkeamman tason $u = 108$ (cm) kohdalla 1910-luvun loppupuolella tapahtuvaa ylitystä seuraava ylitys sattuu vasta 1970-luvun puolivälissä. Jos tasoa vielä nostettaisiin, tapahtuisi ylityksiä vain havaintoaineiston loppupuolella. Lisäksi havaitaan, että ylityksillä on taipumusta sattua lähekkäisinä päivämäärinä. Tämä on tietysti luonnollista, kun merenpinnan taso on korkealla pidemmän aikaa ja havainnot ovat päiväkohtaisia, ja saattaa vaatia perusylitemenetelmää sovellettaessa yliteryppäiden deklusterointia – eli klusterien tunnistamista ja vain maksimihavainnon mukaanottoa kustakin klusterista – jottei datan riippuvuus vääristä tuloksia.

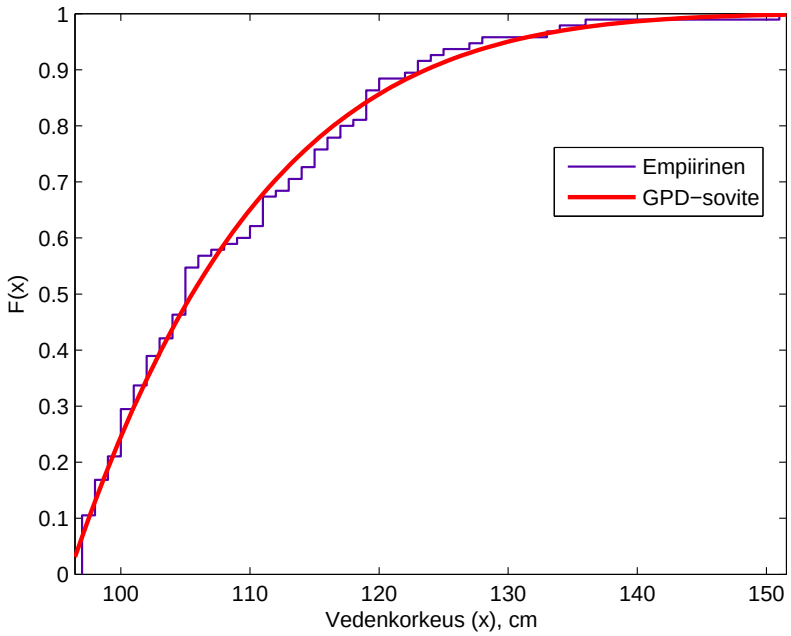
Kuvien mukainen ylitysten sattumisen ajoitus viittaa siihen, ettei ylitedataa voida pitää iid havaintoina. Proposition (1.40) mukaan (ks. myös sitä seuraava perustelu) iid prosessissa korkean kynnyksen ylityksien jakauma noudattaa (approksimatiivisesti) Poisson-jakaumaa. Varsinkaan korkeamman 108 cm:n tason ylitykset eivät kuitenkaan näytä Poisson-jakautuneilta. Tähän palataan myöhemmin, mutta toistaiseksi mahdolliset ongelmat jätetään huomiotta ylitemenetelmän perusmuodon havainnollistamiseksi.

Jatketaan siis sovittamalla yleistetty Pareto-jakauma ylitteisiin suurimman uskottavuuden menetelmällä. Taulukossa 2.7 on esitetty tuloksena saadut parametriestimaatit luottamusväleineen. Erityisesti nähdään, että ylitemenetelmä viittaa vuosimaksimeihin perustuvaa blokkimaksimimenetelmää vahvemmin muotoparametrin ξ negatiivisuuteen; kynnyksen 96 cm tapauksessa nolla ei aivan sisälly 95 %:n luottamusväliin, ja kynnyksen 108 cm kohdalla se sisältyy juuri ja juuri.

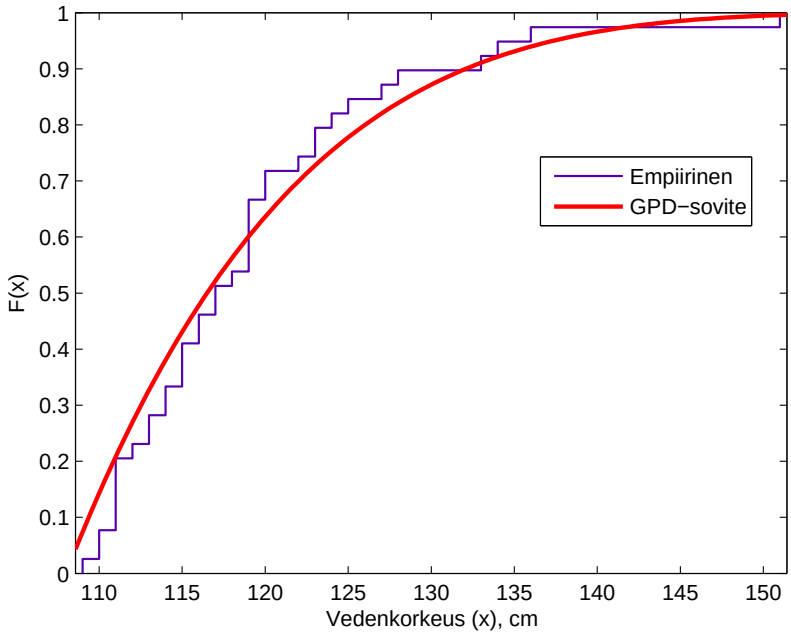
Kuvissa 2.23 ja 2.24 on vertailtu havaintoihin perustuvaa empiiristä kertymäfunktiota vastaavan dataan sovitetun GP-jakauman kertymäfunktion kanssa. Nähdään, että sovite vaikuttaa molemmissa tapauksissa varsin kohtuulliselta.

Taulukko 2.7: Parametriestimaatit GP-jakaumalle valituilla kynnystasoilla.

Kynnys	$u = 96$ (cm)		$u = 108$ (cm)	
Parametri	SUE	95% luottamusväli	SUE	95% luottamusväli
ξ	-0.176	$[-0.341, -0.011]$	-0.208	$[-0.438, 0.023]$
β	14.6	$[11.2, 18.9]$	13.2	$[9.0, 19.3]$



Kuva 2.23: Vedenkorkeuden tason $u = 96$ cm ylittävien havaintojen empiirinen kertymäfunktio vs. GPD-sovite.

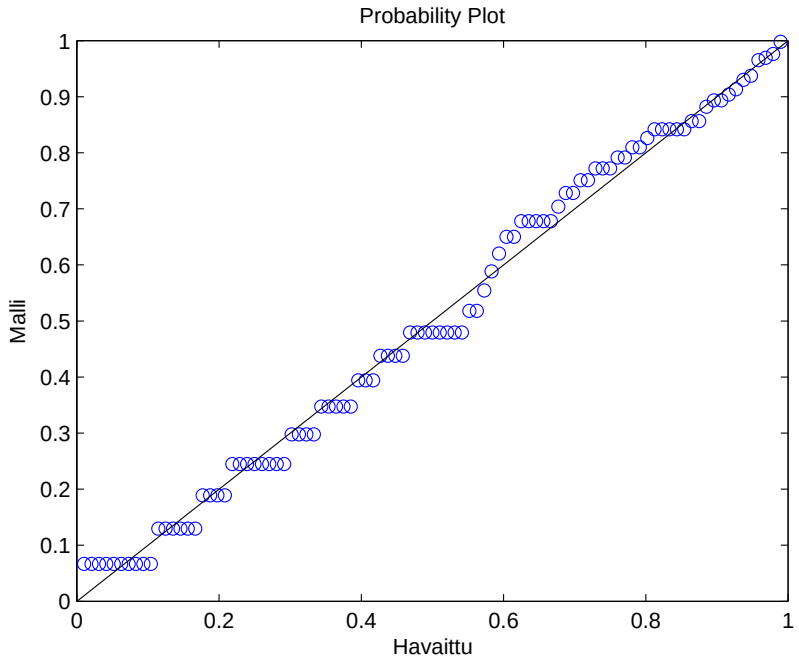


Kuva 2.24: Vedenkorkeuden tason $u = 108$ cm ylittävien havaintojen empiirinen kertymäfunktio vs. GPD-sovite.

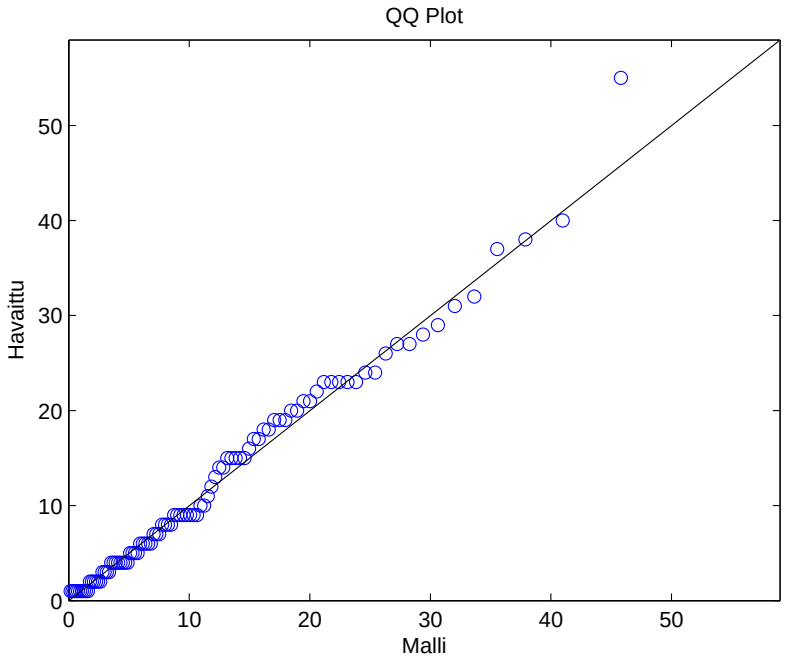
Kuvissa 2.25 ja 2.26 on esitetty todennäköisyys- ja kvantiilikuvaaja tason 96 cm ylitteiden GP-jakaumalle. Kuvissa 2.27 ja 2.28 on vastaavat tason 108 cm ylitteisiin sovitetulle mallille. Myös näiden kuvaajien perusteella mallit sopivat havaintoihin varsin hyvin. Havaintojen diskreetistä luonteesta (mittaukset 1 cm tarkkuudella) johtuen todennäköisyyskuvaajissa nähdään havaintojen horisontaalista kasautumista, kun täsmälleen sama havaintoarvo esiintyy lukuisia kertoja aineistossa (empiiriseen kertymäfunktioon aiheutuu iso porras). Kvantiilikuvaajista erottuu molemmissa tapauksissa suurin vedenkorkeushavainto 151 cm muista "poikkeavana". Suurimman ja toiseksi suurimman havainnon välillä on datan skaalalla huomattava 15 cm:n ero, ja tämä piirre korostuu tarkasteltaessa jakauman häntää.

Tarkastellaan seuraavaksi toistumistasoja. Kuviin 2.29 ja 2.30 on piirretty vuositason toistumistasokuvaajat kynnyksillä $u = 96$ ja $u = 108$ sekä luottamusvälit asymptoottisiin keskivirheisiin perustuen. Lähtödata muodostui päivätason havainnoista, joten N-vuoden toistumistaso vastaa tässä m-havainnon toistumistasoa arvolla $m = n_y N = 365N$ (iid oletuksen pätiessä, tietenkin).

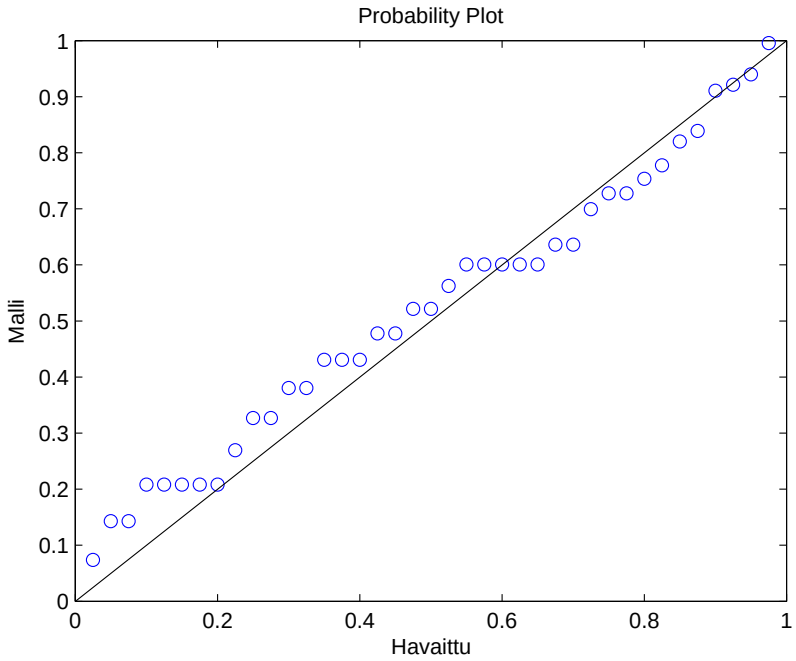
Huomio kiinnittyy kvantiilikuvaajan tavoin suurimpaan, 151 cm:n havaintoon. Tämä ei aivan mahdu keskivirheeseen perustuvien mallin 95 %:n luottamusvälien sisään, toisin kuin muut havainnot. Nähdään myös, että ylitemenetelmän antamat toistumistasot ovat pitkillä aikaväleillä selvästi pienempiä kuin blokkimaksimimenetelmällä saadut. Toisin päin ilmaistuna, korkean merenpinnan tason toistumisperiodi on ylitemenetelmässä paljon pidempi, eli todennäköisyys tason



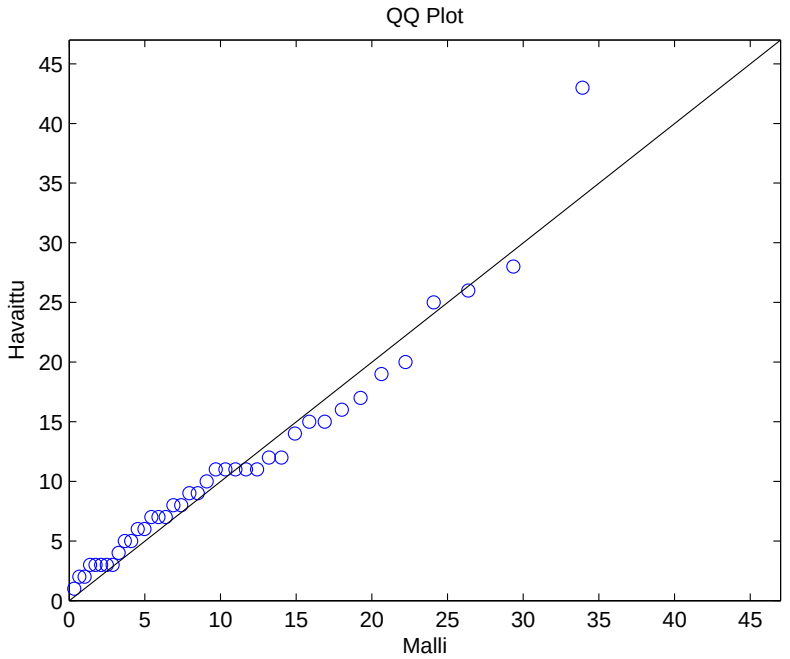
Kuva 2.25: Todennäköisyyskuvaaja tason $u = 96$ cm ylitteiden GPD-sovitteelle.



Kuva 2.26: Kvantiilikuvaaja tason $u = 96$ cm ylitteiden GPD-sovitteelle.

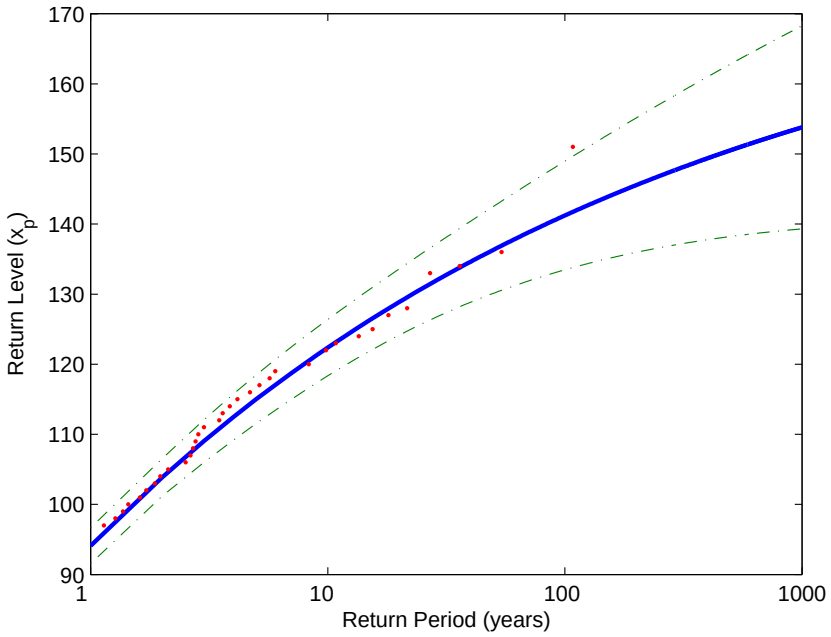


Kuva 2.27: Todennäköisyyskuvaaja tason $u = 108$ cm ylitteiden GPD-sovitteelle.



Kuva 2.28: Kvantiilikuvaaja tason $u = 108$ cm ylitteiden GPD-sovitteelle.

ylittämiseen vuoden aikana pienempi.



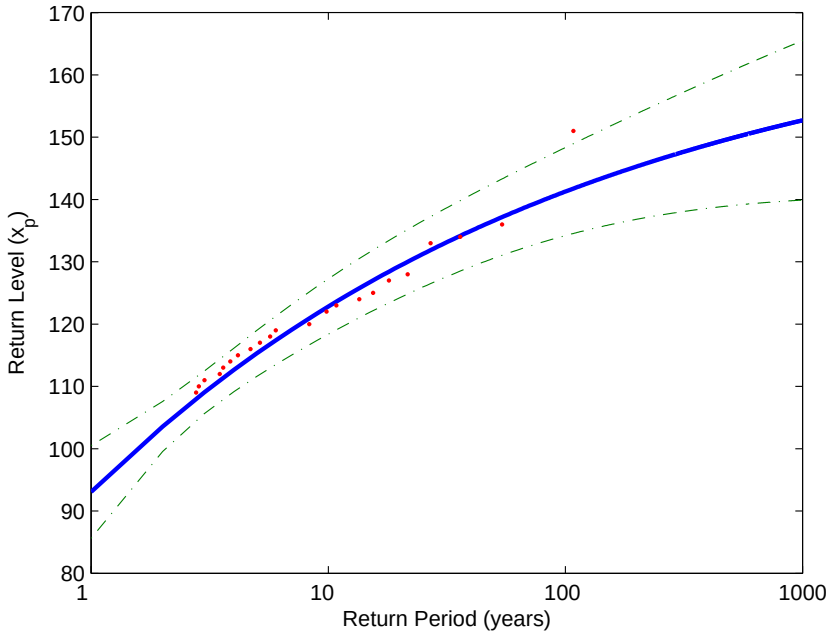
Kuva 2.29: Vedenkorkeuden toistumistasokuvaaja GPD-mallissa ($u = 96$ cm) luottamusväleineen.

SU-estimaattorin asymptoottiseen normaalisuuteen perustuvia luottamusvälejä parempaan tarkkuuteen päästään profiiliuskottavuusmenetelmällä. Alla kuvissa 2.31 ja 2.32 on esitetty profiiliuskottavuusfunktiot 100- ja 1 000-vuoden toistumistasoille kynnyksen $u = 96$ cm ylitteisiin sovitetussa mallissa, ja kuvissa 2.33 ja 2.34 on vastaavat kynnyksen $u = 108$ cm ylitteisiin perustuvassa mallissa. Luottamusvälit saadaan luettua profiiliuskottavuusfunktion ja kriittistä tasoa vastaavan horisontaalisen katkoviivan leikkauspisteistä.

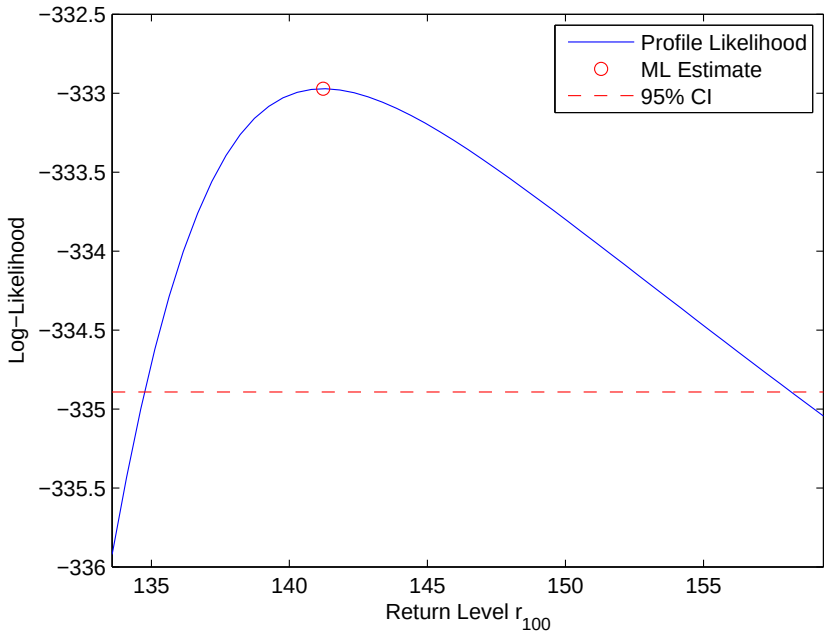
Taulukossa 2.8 on esitetty ja vertailtu eri menetelmien antamia luottamusvälejä eripituisilla toistumisperiodeilla. Profiiliuskottavuus tuo hyvin esille havaintoihin liittyvän epävarmuuden: vaikka esimerkiksi 1 000-vuoden tapahtuman pisteestimaatti on ”vain” n. 153 cm – ja sellaisenaan epäuskottavan alhainen[12] – ulottuu 95 %:n luottamusväli sille 144 senttimetristä 193 tai 198 senttimetriin.

Tarkastellaan seuraavaksi samaa asiaa hieman eri näkökulmasta, eli vedenkorkeustasojen ylitystodennäköisyyksiä. Todennäköisyyksia tarkastellessa täytyy muistaa, että sovitetty GPD-malli kuvaa tason u ylitejakaumaa, eli ylitteiden jakaumaa *ehdolla*, että taso u ylitetään. GP-jakauman suoraan antamat todennäköisyydet ovat siis ehdollisia todennäköisyyksiä (vrt. kertymäfunktioiden kuvaajat edellä). Kuten toistumisperiodien yhteydessä osiossa 2.3.3 nähtiin, ehdollistamaton ylitteen – eli tapahtuman $X > x$, kun $x > u$ – todennäköisyys

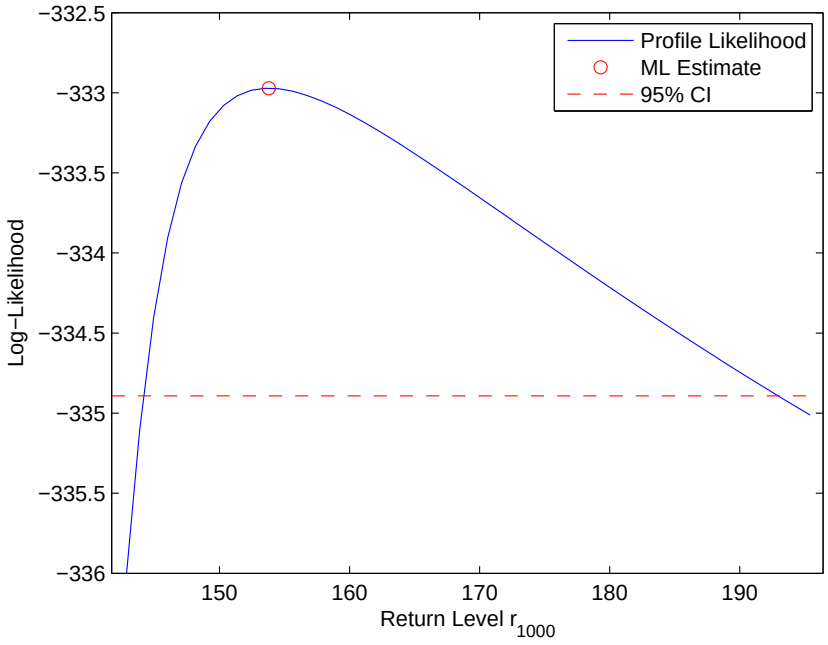
[12] Vrt. keskustelu tämän osion lopussa.



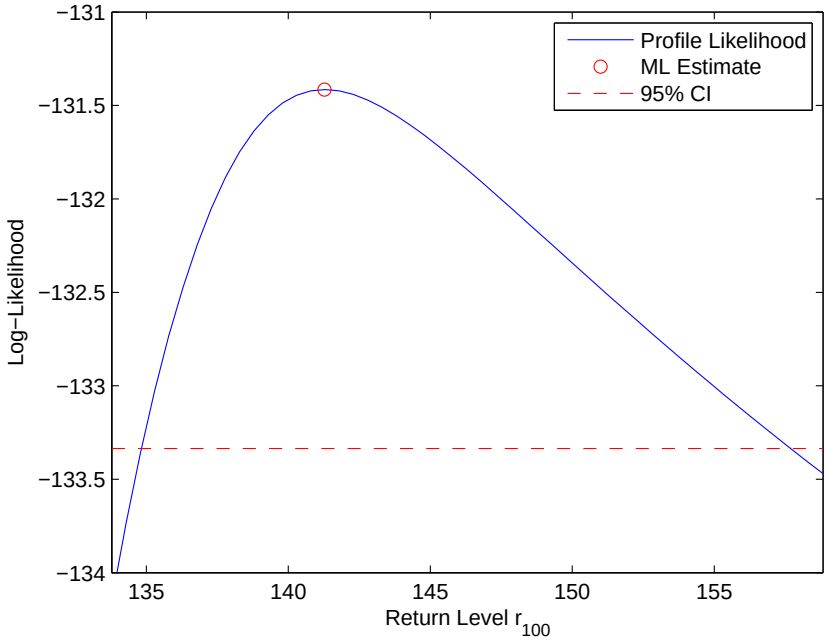
Kuva 2.30: Vedenkorkeuden toistumistasokuvaaja GPD-mallissa ($u = 108$ cm) luottamusväleineen.



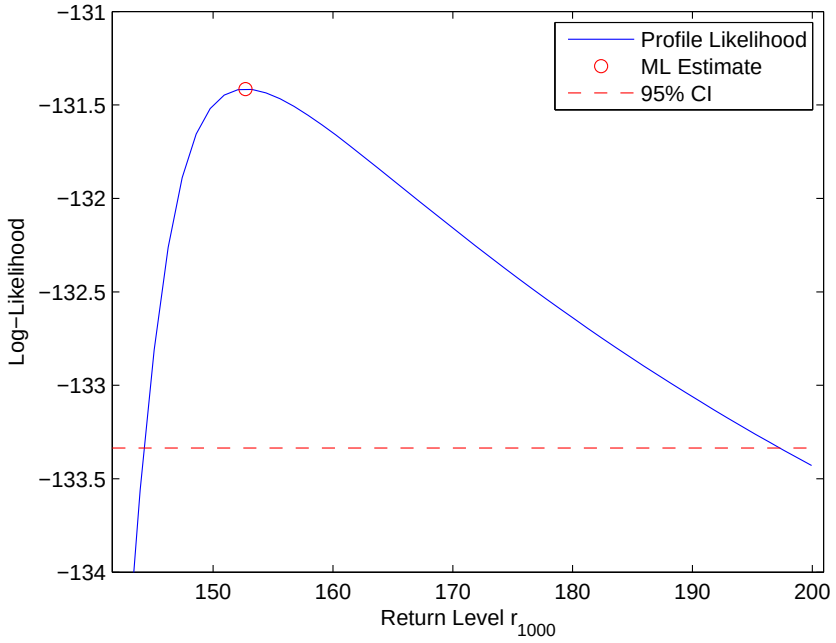
Kuva 2.31: Profiiliuskottavuus merenpinnan korkeuden 100-vuoden toistumistasolle GPD-mallissa ($u = 96$ cm).



Kuva 2.32: Profiiliuskottavuus merenpinnan korkeuden 1 000-vuoden toistumistasolle GPD-mallissa ($u = 96$ cm).



Kuva 2.33: Profiiliuskottavuus merenpinnan korkeuden 100-vuoden toistumistasolle GPD-mallissa ($u = 108$ cm).



Kuva 2.34: Profiiliuskottavuus merenpinnan korkeuden 1 000-vuoden toistumistasolle GPD-mallissa ($u = 108$ cm).

Taulukko 2.8: Delta-menetelmään ja profiliuskottavuuteen perustuvat 95 %:n luottamusvälit merenpinnan korkeuden toistumistasoille GPD-malleissa.

$u = 96$	Delta-menetelmä		Profiiliuskottavuus	
Toistumisperiodi	SUE (cm)	95% CI (cm)	SUE (cm)	95% CI (cm)
10	122	[118, 127]	122	[118, 128]
100	141	[133, 149]	141	[134, 159]
1 000	154	[139, 167]	154	[144, 193]
$u = 108$	Delta-menetelmä		Profiiliuskottavuus	
Toistumisperiodi	SUE (cm)	95% CI (cm)	SUE (cm)	95% CI (cm)
10	122	[118, 128]	122	[119, 128]
100	141	[134, 149]	141	[135, 158]
1 000	153	[139, 166]	153	[144, 198]

voidaan (osion merkintöjä käyttäen) kirjoittaa

$$\mathbb{P}(X > x) = \mathbb{P}(X > u)\mathbb{P}(X > x|X > u) = \zeta_u \bar{G}_{\boldsymbol{\theta}}(x - u) =: 1 - p.$$

Toisin sanoen todennäköisyys, että havainto ei ylitä tasoa x on $\mathbb{P}(X \leq x) = p$. Edelleen todennäköisyys, että tasoa x ei ylitetä vuodessa on yhtä kuin todennäköisyys, että tasoa ei ylitetä minään päivänä vuoden kuluessa. Kun merkitään $\{X^v > x\}$ tapahtumaa, että x ylitetään vuoden aikana, saadaan tapahtuman todennäköisyydeksi

$$\begin{aligned}\mathbb{P}(X^v > x) &= \mathbb{P}(X_i > x \text{ jollakin } i = 1, \ldots, 365) \\ &= 1 - \mathbb{P}(X_i \leq x \; \forall \, i = 1, \ldots, 365) \\ &= 1 - \mathbb{P}(X \leq x)^{365} \\ &= 1 - p^{365},\end{aligned}$$

missä X_i, $i = 1, \ldots, 365$ ovat päivittäisiä havaintoja, $X_i \sim X \; \forall \, i$, ja havaintojen oletetaan olevan riippumattomia. Sama tulos saadaan myös suoraan binomijakauman ominaisuuksiin perustuen, huomaamalla että ylitteiden lukumäärä on binomijakautunut satunnaismuuttuja.

Taulukkoon 2.9 on koottu edellä kuvatulla tavalla laskettuja vedenkorkeuden rajatasoja x_p eri ylitystodennäköisyyksille $1 - p$.[13] Verrattuna blokkimaksimimenetelmän (ks. taulukko 2.4) antamiin vedenkorkeuksia vastaaviin ylitystodennäköisyyksiin, ylitemenetelmän mukaan keskikorkeiden tasojen ylittämisen todennäköisyys on suurempi, mutta hyvin korkeiden tasojen ylittämisen todennäköisyys selvästi pienempi. Ero kasvaa mitä pidemmälle vedenkorkeusjakauman häntään mennään.

Taulukko 2.9: Eri todennäköisyyksiä vastaavia merenpinnan korkeustasoja GPD-malleissa.

p	$1 - p$	x_p, $u = 96$	x_p, $u = 108$
0.5	0.5	99 (cm)	-
0.75	0.25	111 (cm)	111 (cm)
0.90	0.10	122 (cm)	122 (cm)
0.99	0.01	141 (cm)	141 (cm)
0.999	0.001	154 (cm)	153 (cm)

Suoraviivaisesti sovelletun ylitemenetelmän mukaan siis 2005 mitattu 151 cm merenpinnan korkeus Helsingissä olisi todellakin äärimmäinen havainto. Oletus ylitteiden riippumattomuudesta ja samoin jakautuneisuudesta (iid-oletus) ei kuitenkaan näytä kuvien 2.21 ja 2.21 perusteella pätevän, mikä saattaa ylitemenetelmällä – sen tässä osiossa käytetyssä perusmuodossa – saatujen tuloksien tarkkuuden kyseenalaiseksi. Lisätarkastelu on selvästi tarpeen. Tähän palataan osiossa 2.6 pisteprosessien yhteydessä.

[13] Todennäköisyyttä 0.5 vastaava korkeus x_p tason $u = 108$ cm ylitteisiin sovitetulle GPD-mallille on taulukossa tyhjä, koska ylitejakauma $F_u(x) = G_{\boldsymbol{\theta}}(x)$ on määritelty vain, kun $x \geq u$ $(= 108)$.

2.4 Stationaariset aikasarjat

Kappaleessa 1.4 kävi ilmi, että stationaarisiin aikasarjoihin on mahdollista soveltaa pitkälti samoja perustekniikoita kuin niiden riippumattomiin ja samoin jakautuneisiin vastineisiinkin. Tarkastellaan seuraavaksi lyhyesti, millaisia lisähuomiota vaativia seikkoja stationaarisuus tuo ääriarvojen tilastolliseen mallintamiseen.

Osiosta 1.4 muistetaan ehdot $D(u_n)$ ja $D'(u_n)$, jotka yhdessä varmistavat, että ehdot täyttävän stationaarisen prosessin ääriarvoilla on sama asymptoottinen käyttäytyminen kuin vastaavalla iid jonolla.

2.4.1 Blokkimaksimimenetelmä

Mikäli stationaarisen prosessi pitkän kantaman riippuvuus prosessin korkeilla tasoilla on heikkoa (täsmällisesti, prosessi täyttää ehdon $D(u_n)$), voidaan maksimien jakaumaa mallintaa GEV-jakaumalla kuten iid tapauksessakin. Mikäli ehto $D'(u_n)$ ei täyty, eli prosessin korkeilla arvoilla on taipumusta esiintyä klustereissa, tarvitaan ääriarvoindeksin käsitettä (määritelmä 1.41): intuitiivisesti, ääriarvoindeksi kuvaa prosessin ääriarvojen klusteroitumistaipumuksen vahvuutta. Kun $\theta < 1$, ääriarvojen ryppäinä esiintyminen vaikuttaa rajajakauman parametrien arvoihin ääriarvoindeksin kautta; klusteroituminen on sitä voimakkaampaa, mitä lähempänä θ on nollaa. Jos $\theta = 1$, prosessilla ei ole taipumusta kasaantua suurilla arvoilla ja maksimit käyttäytyvät täsmälleen kuten iid-tapauksessakin (jolle aina $\theta \equiv 1$).

Käytännössä ääriarvoindeksin $\theta < 1$ vaikutus uppoaa kokonaan GEV-jakauman parametreihin. Johtopäätöksenä on siis, että blokkimaksimimenetelmän yhteydessä stationaarisen aikasarjan ilmentämä riippuvuus voidaan mallinnusmielessä jättää huomiotta, ja stationaariseen tapaukseen soveltaa täsmälleen samoja tilastomenetelmiä kuin iid tapaukseenkin. Riippuvuus kuitenkin todennäköisesti vaikuttaa tilastomenetelmien perusteluna olevan GEV-approksimaation tarkkuuteen heikentävästi, vaikka tarkka asymptoottinen rajatulos onkin (ääriarvoindeksin vaikutusta vaille) täsmälleen sama iid- ja stationaariselle jonolle.

2.4.2 Ylitemenetelmä

Kuten GEV-jakauma blokkimaksimimenetelmässä, GP-jakauma säilyy ylitteille soveltuvana stationaarisenkin sarjan kohdalla. Jos $\theta = 1$, niin muutoksia iid tapaukseen ei tilastollisten menetelmien osalta tarvita. Jos taas $\theta < 1$ eli prosessi osoittaa taipumusta ääriarvojen klusteroitumiseen, täytyy tämä huomioida mallinnuksessa. Blokkimaksimimenetelmän kohdalla ongelma olennaisesti kierrettiin tarkastelemalla vain n-blokkien maksimihavaintoja, missä taustaoletuksena on että blokit ovat riittävän suuria. Sen sijaan ylitemenetelmässä klusteroituminen voi aiheuttaa useita vierekkäisiä valitun korkean tason ylityksiä: Tilastollisen mallinnuksen perusteluna olevat asymptoottiset tulokset kertovat, että korkean kynnyksen ylittävien havaintojen reunajakauma on yleistetty Pareto, mutta eivät määrää vierekkäisten ylitysten yhteisjakaumaa. Tämä tarkoittaa, että (riippumattomuuteen perustuva) log-uskottavuus (2.11) ei päde

tässä tapauksessa enää edes approksimaationa. Yleistä teoriaa, joka tarjoaisi riippuvuuden huomioonottavan uskottavuusfunktion muodon, ei myöskään ole olemassa.

Käytännössä ongelmaa lähestytään usein pyrkimällä tunnistamaan erilliset havaintoklusterit, ja ottamalla mallinnukseen mukaan vain kunkin klusterin suurin havainto, klusterimaksimi. Mikäli klusterit on tunnistettu oikein, pitäisi näin saatujen havaintojen olla (likimain) riippumattomia, jolloin tavanomaisia mallinnustekniikoita voidaan taas soveltaa. Menetelmää kutsutaan deklusteroinniksi.

Yksinkertaisimmallaan deklusterointi voi koostua havaintojen manuaalisesta tutkimisesta ja klusterimaksimien valinnasta, mutta myös automaattisia tekniikoita on kehitetty. Yksinkertainen tapa määritellä klusteri on määrittää kynnystaso u, ja pitää vierekkäisiä kynnyksen ylityksiä samaan ryppääseen kuuluvina. Tason u alitus lopettaa tällöin klusterin, ja seuraavasta ylityksestä alkaa seuraava klusteri. Yleensä on tarpeen sallia useampia alituksia ennen kuin klusterin katsotaan katkenneen, jolloin ylitteiden katsotaan kuuluvan samaan klusteriin, kunnes r peräkkäistä havaintoa putoaa kynnyksen u alle, missä $r \in \mathbb{N}$ on etukäteen kiinnitetty luku. Lukuarvon r valinnalla voi olla suuri vaikutus: liian pieni r tarkoittaa, että tuloksena saatavia vierekkäisiä klustereita ei todennäköisesti voi pitää riippumattomina, ja liian suuri r puolestaan, että riippumattomat klusterit sulautetaankin yhteen.

Deklusterointi tässä yksinkertaisessa muodossaan on menetelmänä yksinkertainen, mutta luonteeltaan *ad hoc*. Menetelmän haittapuolia on erityisesti se, että tulokset voivat olla herkkiä deklusterointisäännön (esim. lukuarvon r) valinnalle. Menetelmä myös jättää huomiotta mahdollisesti paljonkin dataa, kun vain kunkin klusterin maksimihavainto huomioidaan. Perustavanlaatuisemmin, deklusteroinnilla pyritään poistamaan lyhyen kantaman riippuvuus ja filtteröimään havainnoista riippumattomia. Useissa sovelluksissa – esimerkiksi markkinariskin kohdalla, ks. luku 4 – kuitenkin juuri tämän riippuvuuden mallintaminen on kiinnostavaa ja tarpeellista.

2.5 Epästationaariset aikasarjat

Monissa ilmiöissä on havaittavissa piirteitä jotka muuttuvat systemaattisesti ajassa. Esimerkiksi sääilmiöihin liittyvät aikasarjat voivat ilmentää kausivaihtelua eli syklejä eri vuodenaikoina vallitsevista erilaisista ilmasto-olosuhteista johtuen, tai niissä voi olla havaittavissa trendi esimerkiksi ilmaston muutoksen tai maankohoamisen seurauksena. Tällaisia prosesseja kutsutaan epästationaarisiksi.

Epästationaarisiin aikasarjoihin ei voida suoraan soveltaa iid-oletukseen perustuvia perusmuotoisia ääriarvomalleja. Osioissa 1.4 ja 2.4 nähtiin, että stationaaristen prosessien kohdalla klassisen ääriarvoteorian malleja voidaan – tietyin rajoituksin ja tiettyjen ehtojen voimassaollessa – pitkälti soveltaa stationaarisiinkin sarjoihin niiden ilmentämästä riippuvuudesta huolimatta. Epästationaaristen prosessien kohdalla vastaavaa yleistä teoriaa ei ole olemassa, vaan käytännössä epästationaarisuus täytyy huomioida tilannekohtaisesti asianmukaisella tilastollisella mallinnuksella. Tällöin ääriarvoteorian perusmallit – blokkimaksimi-

menetelmän tai ylitemenetelmän mukaiset – otetaan mallinnuksen pohjaksi, ja niitä laajennetaan ottamaan huomioon ilmiöiden epästationaarisuus.

Epästationaarisuus havaintoaikasarjoissa on luonnollisinta ottaa huomioon asettamalla GEV- tai GP-jakauman parametrit riippumaan ajasta tai muista selittävistä muuttujista (covariates). Blokkimaksimimenetelmän tapauksessa tämä tarkoittaa, että (vuosi)maksimien (X_t) oletetaan noudattavan GEV-jakaumaa

$$X_t \sim H_{\boldsymbol{\theta}(t)} = H(\boldsymbol{\xi}(t), \boldsymbol{\mu}(t), \boldsymbol{\sigma}(t))$$

hetkellä t. Vastaavasti GP-jakaumaan perustuvien mallien kohdalla ylitemenetelmässä taustaoletuksena on, että ylitteet noudattavat jakaumaa

$$(X_t - u(t) | X_t > u(t)) \sim G_{\boldsymbol{\theta}(t)} = G(\boldsymbol{\xi}(t), \boldsymbol{\beta}(t)),$$

missä $u(t)$ on mahdollisesti ajasta riippuva kynnystaso. GP-jakauman kohdalla epästationaarisuutta on helpompaa – ja luonnollisempaa – käsitellä pisteprosessien kontekstissa, kuten myöhemmin tullaan näkemään.

Esimerkiksi lineaarinen trendi lokaatioparametrissa voidaan lisätä GEV-malliin asettamalla $X_t \sim H(\xi, \boldsymbol{\mu}(t), \sigma)$, missä

$$\boldsymbol{\mu}(t) = \kappa_0 + \kappa_1 t,$$

ja mallin parametrivektoriksi tulee $(\xi, \kappa_0, \kappa_1, \sigma)$. Vastaavasti jos tarkasteltavassa aikasarjassa on havaittavissa tason muutos (regime change), voidaan tätä mallintaa asettamalla esimerkiksi

$$\boldsymbol{\mu}(t) = \begin{cases} \kappa_0, & \text{kun} \quad t < t_0, \\ \kappa_1, & \text{kun} \quad t \geq t_0, \end{cases}$$

missä t_0 on muutospiste. Kausivaihtelu voidaan pyrkiä kuvaamaan asettamalla parametrit riippumaan kaudesta, esimerkiksi kuukaudesta vuoden sisällä (yleensä paloittain vakioina tai paloittain lineaarisina), tai käyttämällä periodisia funktioita syklin kuvaamiseen:

$$\boldsymbol{\mu}(t) = \kappa_0 + \kappa_1 \sin\left(\frac{2\pi t}{T}\right) + \kappa_2 \sin\left(\frac{2\pi t}{T}\right),$$

missä T on syklin pituus. Ulkoiset muuttujat – kuten vaikkapa maastopaloihin vaikuttavat kuivuusolosuhteet – voidaan lisätä malliin samaan tapaan:

$$\boldsymbol{\mu}(t) = \kappa_0 + \kappa_1 W(t),$$

missä selittävän muuttujan W arvo hetkellä t on $W(t)$.

Vastaavia malleja voidaan soveltaa myös skaalaparametriin σ (tai β) muistaen kuitenkin pitää huolta skaalaparametrin positiivisuudesta, esimerkiksi mallintamalla log-muunnoksen $\ln \sigma$ käyttäytymistä. Sen sijaan muotoparametria ξ (tai vastaavaa häntäindeksia $\alpha = 1/\xi$ stabiileilla jakaumilla) on pahamaineisen vaikeaa estimoida datasta tarkasti [2]; ξ:n mallintaminen sileänä (smooth) funktiona ajan suhteen onkin yleensä epärealistinen tavoite. Kausivaihtelua kuvaavassa mallissa saattaa syntyä tarve määrittää ξ paloittain vakioksi, vastaten ilmiön

eri kausina ilmentämää erilaista luonnetta. Useimmissa sovelluksissa muuttuvalle muotoparametrille ei kuitenkaan löydy perusteita, ja se pidetäänkin yleensä vakiona.

Kaikki edellä esitetyt esimerkit voidaan kirjoittaa yleisessä muodossa

$$\boldsymbol{\theta}(t) = \begin{pmatrix} \boldsymbol{\xi}(t) \\ \boldsymbol{\mu}(t) \\ \boldsymbol{\sigma}(t) \end{pmatrix} = \begin{pmatrix} h_\xi(\boldsymbol{X}_\xi^T \boldsymbol{\kappa}_\xi) \\ h_\mu(\boldsymbol{X}_\mu^T \boldsymbol{\kappa}_\mu) \\ h_\sigma(\boldsymbol{X}_\sigma^T \boldsymbol{\kappa}_\sigma) \end{pmatrix},$$

missä $\boldsymbol{\kappa}_\theta$ on jakaumaparametrin θ mallia vastaava parametrien vektori, $\boldsymbol{X}_\theta$ vastaava mallivektori (design vector), ja $h_\theta = \eta_\theta^{-1}$ sopiva funktio; funktiota η kutsutaan yleensä linkkifunktioksi, jolloin h on käänteinen linkkifunktio (inverse link function). Asetelma on tuttu yleistettyjen lineaaristen mallien (GLM) yhteydestä. GLM:iä koskeva teoria tai niitä varten kehitetyt tilastolliset työkalut eivät kuitenkaan suoraan sovi ääriarvo-ongelmiin, mm. koska yleistetyt lineaariset mallit rajoittuvat jakaumiin, jotka kuuluvat eksponenttijakaumaperheeseen (exponential family of distributions). Ääriarvojakaumat eivät yleisesti kuulu edellä mainittuun.

Esimerkiksi yllä lineaarisen trendin tapauksessa

$$\boldsymbol{\mu}(t) = h(\boldsymbol{X}^T \boldsymbol{\kappa}) = (1, t) \begin{pmatrix} \kappa_0 \\ \kappa_1 \end{pmatrix}$$

identiteettilinkillä $\eta(x) \equiv 1 \equiv h(x)$, ja muutospistemallissa kahdella muutospisteellä

$$\boldsymbol{\mu}(t) = h(\boldsymbol{X}^T \boldsymbol{\kappa}) = \left(\mathbb{1}_{\{t<t_0\}}, \mathbb{1}_{\{t_0 \le t<t_1\}}, \mathbb{1}_{\{t>t_1\}}\right) \begin{pmatrix} \kappa_0 \\ \kappa_1 \\ \kappa_2 \end{pmatrix}$$

niin ikään identiteettilinkillä. Kuten yllä mainittiin, skaalaparametrin σ kohdalla käytetään usein logaritmistä linkkifunktiota eli log-linkkiä $\eta(x) = \ln x$, jolloin h on eksponenttifunktio, $h(x) = \eta^{-1}(x) = \exp(x)$.

2.5.1 Suurimman uskottavuuden menetelmä epästationaarisille prosesseille

Yksi suurimman uskottavuuden menetelmän vahvuuksista on sen helppo muokattavuus monimutkaisillekin mallirakenteille. Tarkastellaan epästationaarista GEV-mallia jonolle $(X_t)_{t=1}^n$:

$$X_t \sim H(\boldsymbol{\theta}(t)),$$

missä $\boldsymbol{\theta}(t) = (\boldsymbol{\xi}(t), \boldsymbol{\mu}(t), \boldsymbol{\sigma}(t))$, ja merkitään havaittua otosta $\boldsymbol{X} = \boldsymbol{x} = (x_1, \ldots, x_n)$. Uskottavuusfunktio on nyt

$$L(\boldsymbol{\kappa}; \boldsymbol{x}) = \prod_{t=1}^{n} h(x_t; \boldsymbol{\xi}(t), \boldsymbol{\mu}(t), \boldsymbol{\sigma}(t)),$$

missä $\boldsymbol{\kappa} = (\boldsymbol{\kappa}_\xi, \boldsymbol{\kappa}_\mu, \boldsymbol{\kappa}_\sigma)$. Log-uskottavuusfunktioksi auki kirjoitettuna tulee

$$l(\boldsymbol{\kappa}; \boldsymbol{x}) = -\sum_{t=1}^{n} \left\{ \ln \boldsymbol{\sigma}(t) + \left(1 + \frac{1}{\boldsymbol{\xi}(t)}\right) \ln \left(1 + \boldsymbol{\xi}(t) \frac{x_t - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)}\right) + \left(1 + \boldsymbol{\xi}(t) \frac{x_t - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)}\right)^{-1/\boldsymbol{\xi}(t)} \right\}, \tag{2.17}$$

kun $1 + \boldsymbol{\xi}(t)(x_t - \boldsymbol{\mu}(t))/\boldsymbol{\sigma}(t) > 0$, kaikilla $t \in I = 1, \ldots, n$. Mikäli $\xi(t) = 0$ joillakin indekseillä $t \in I_0$, $I_0 \subset I$, niin näitä vastaavat termit yo. summassa tulee korvata log-uskottavuusfunktion Gumbel-muodolla,

$$l(\boldsymbol{\kappa}; \boldsymbol{x}) = -\sum_{t=1}^{n} \left\{ \ln \boldsymbol{\sigma}(t) + \frac{x_t - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)} + \exp\left(-\frac{x_t - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)}\right) \right\}. \tag{2.18}$$

Ts. jos $\xi(t_0) = 0$, niin aikaindeksiä t_0 vastaavan termin kontribuutio log-uskottavuuteen on

$$l(\boldsymbol{\kappa}; x_{t_0}) = -\left\{ \ln \boldsymbol{\sigma}(t_0) + \frac{x_{t_0} - \boldsymbol{\mu}(t_0)}{\boldsymbol{\sigma}(t_0)} + \exp\left(-\frac{x_{t_0} - \boldsymbol{\mu}(t_0)}{\boldsymbol{\sigma}(t_0)}\right) \right\}.$$

Yllä esitetyissä lausekkeissa tulee $\boldsymbol{\xi}(t)$, $\boldsymbol{\mu}(t)$ ja $\boldsymbol{\sigma}(t)$ korvata niille käytetyillä malleilla, esim. $\boldsymbol{\mu}(t) = \kappa_\mu^0 + \kappa_\mu^1 t$, jolloin saadaan parametrisaatio malliparametrien $\boldsymbol{\kappa}$ suhteen.

2.5.2 Mallin valinta

Ääriarvomallien parametrien mallintaminen ajan ja selittävien muuttujien funktiona johtaa suureen määrään potentiaalisia mallirakenteita. Tällöin kysymykseksi nousee, miten valita ehdokasmalleista paras? Ilmiöiden fysikaalinen luonne, sovellusaluekohtainen taustateoria ja näihin perustuvat järkevyystarkastelut usein ohjaavat valintaa ja rajoittavat potentiaalisten mallien luokkaa, mutta tämänkin jälkeen on yleensä tarjolla lukuisa määrä mahdollisia mallirakenteita. Perusperiaatteena kaikessa tilastollisessa mallinnuksessa on pyrkimys parsimoniaan eli kuvaamaan dataa yksinkertaisimmalla mahdollisella mallilla, joka selittää datan sisältämästä vaihtelusta niin paljon kuin mahdollista (*principle of parsimony*). Tällöin malleja vertailtaessa tarkastellaan, kasvattaako ylimääräisen parametrin lisääminen malliin mallin selitysvoimaa "riittävästi" (tilastollisesti merkitsevästi), jotta mallin monimutkaistaminen olisi perusteltua.

Kun mallien estimoinnissa käytetään suurimman uskottavuuden menetelmää, voidaan sisäkkäisiä mallirakenteita verrata helposti toisiinsa uskottavuusosamäärätestin avulla. Merkitään verrattavia malleja $\mathcal{M}_0$ ja $\mathcal{M}_1$, missä $\mathcal{M}_0 \subset \mathcal{M}_1$, eli malli $\mathcal{M}_0$ on yleisemmän mallin erikoistapaus. Tällöin mallinvalinta voidaan perustaa testisuureeseen (devianssiin)

$$D = 2\left(l_1(\mathcal{M}_1) - l_0(\mathcal{M}_0)\right),$$

missä $l_1(\mathcal{M}_1)$ ja $l_0(\mathcal{M}_0)$ ovat malleja vastaavat maksimoidut log-uskottavuudet. Suuret testisuureen D arvot viittaavat siihen, että malli $\mathcal{M}_1$ selittää merkittävästi enemmän datan sisältämästä vaihtelusta, kun taas pienet arvot viittaavat siihen, että mallin koon kasvattaminen ei paranna mallin selityskykyä merkittävästi. Täsmällisemmin, malli $\mathcal{M}_0$ hylätään merkittävyystasolla α, jos $D >$

c_α, missä kriittinen taso c_α on χ^2_k-jakauman $(1-\alpha)$-kvantiili, kun mallien $\mathcal{M}_1$ ja $\mathcal{M}_0$ dimensioiden ero on k; ks. tarkemmin liite B.

Jos mallit eivät ole sisäkkäisiä, ei niitä voi suoraan verrata keskenään uskottavuusosamäärätestiin perustuen samaan tapaan kuin yllä. Ei-sisäkkäisten mallien vertailuun on ehdotettu useita tunnuslukuja, ns. informaatiokriteereitä, joiden perusajatuksena on ottaa lähtökohdaksi mallilla saavutettava maksimaalinen log-uskottavuus, ja vähentää tästä malliparametrien määrään verrannollinen sakko. Pyrkimyksenä on siis saattaa eri mallirakenteet vertailukelpoisiksi huomioimalla niiden parametrien lukumäärä; ks. tarkemmin liite B. Yleisesti käytettyjä informaatiokriteerejä ovat mm. Akaiken informaatiokriteeri (AIC) ja Bayesilainen informaatiokriteeri (BIC).

2.5.3 Toistumisperiodi ja toistumistaso

Epästationaaristen mallien tapauksessa toistumistason määrittely ei ole yksikäsitteistä, sillä prosessin taustalla oleva todennäköisyysjakauma muuttuu ajan suhteen. Tällöin ei voida puhua "t-vuoden tapahtumasta" tai kerran t:ssä vuodessa ylitettävästä tasosta – kuten iid- tai stationaarisen tapauksen yhteydessä – täsmentämättä, mitä tällä tarkoitetaan.

Epästationaarisessa tapauksessa toistumistason täsmällinen laskutapa riippuu käytetyn mallin muodosta. Esimerkiksi n-havainnon tapauksessa m:n n-blokin toistumistaso x_m voidaan määritellä yhtälöstä

$$\frac{1}{m} = \mathbb{P}(\max(X_1, \ldots, X_n) > x_m) = 1 - \mathbb{P}(\max(X_1, \ldots, X_n) \leq x_m)$$

Jos havainnot (X_n) oletetaan riippumattomiksi ehdolla $\boldsymbol{\theta}$, saadaan edelleen

$$\frac{1}{m} = 1 - \prod_{t=1}^{n} F_{\boldsymbol{\theta}(t)}(x_m),$$

missä $F_{\boldsymbol{\theta}(t)} = H_{\boldsymbol{\theta}(t)}$ tai $F_{\boldsymbol{\theta}(t)} = G_{\boldsymbol{\theta}(t)}$. Toistumistason estimaatti saadaan ratkaisemalla yhtälöstä x_m estimoiduilla parametrien arvoilla $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

Esimerkiksi mallissa, jossa on vuoden mittainen sykli ja n havaintoa vuodessa, yllä oleva vastaa suoraan m-vuoden toistumistasoa. Jos havainnot $1, \ldots, n$ sen sijaan ovat vuosittaisia, vastaa toistumistaso x_m m:n n:n vuoden blokin eli mn-vuoden toistumistasoa. Siten "keskimääräinen" tai efektiivinen m-vuoden toistumistaso saadaan jakamalla vuosien lukumäärällä eli n:llä. Kun mallissa on esimerkiksi trendi, täytyy lisäksi määritellä, mitä n-vuoden mittaista ajanjaksoa tarkoitetaan, sillä tulos riippuu tarkasteltavasta ajanjaksosta (todennäköisyyksien tulo vaikkapa väleillä $t = 1, \ldots, n$ ja $t = 2, \ldots, n+1$ on erilainen). Toistumistaso ja toistumisperiodi -terminologia onkin käyttökelpoisinta, kun havainnot ovat samoin jakautuneita (eli iid tai stationaarisia). Epästationaarisessa tapauksessa näiden käsitteiden käyttö puolestaan ei aina ole kovin luonnollista.[14]

[14] Ks. kohta 2.6.3.3 konkreettista esimerkkiä varten, jossa *toistumis*tason sijaan tarkastellaan odotusarvoisesti kerran *seuraavassa* m vuodessa sattuvaa tapahtumaa.

2.5.4 Mallidiagnostiikkaa

Epästationaarisissa malleissa kutakin otoksen havaintoa voi vastata erilainen todennäköisyysjakauma. Mallin hyvyyden arvioimiseksi todennäköisyys- ja kvantiilikuvaajien avulla täytyy havainnot ensin saattaa noudattamaan samaa jakaumaa standardisoimalla data.

Estimoidun epästationaarisen GEV-mallin,

$$X_t \sim H_{\boldsymbol{\theta}(t)} = H(\boldsymbol{\xi}(t), \boldsymbol{\mu}(t), \boldsymbol{\sigma}(t)),$$

kohdalla esimerkiksi residuaalit

$$\varepsilon_t = \bar{X}_t = \left(1 + \boldsymbol{\xi}(t)\frac{X_t - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)}\right)^{-1/\boldsymbol{\xi}(t)}$$

noudattavat standardia eksponenttijakaumaa, $\mathbb{P}(\bar{X}_t \leq x) = 1 - \exp(-x)$, $x \in \mathbb{R}$, ja residuaalit

$$\varepsilon_t = \bar{X}_t = \frac{1}{\boldsymbol{\xi}(t)} \ln\left(1 + \boldsymbol{\xi}(t)\frac{X_t - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)}\right)$$

standardia Gumbel-jakaumaa, $\mathbb{P}(\bar{X}_t \leq x) = \exp(-e^{-x})$, $x \in \mathbb{R}$. Todennäköisyyskuvaaja on invariantti referenssijakauman valinnan suhteen, mutta kvantiilikuvaaja ei. Tavanomainen valinta referenssijakaumaksi on kuitenkin Gumbel, johtuen mm. sen luonnollisesta asemasta GEV-jakaumaperheessä.

Oletetaan että havaintojen muunnos tehdään Gumbel-jakaumaan, ja merkitään havaittua otosta $\boldsymbol{X} = \boldsymbol{x} = (x_1, \ldots, x_n)$, jolloin havaitut residuaalit ovat $\bar{x}_t$, $t = 1, \ldots, n$, ja näiden järjestetty otos $\bar{x}_{(1)}, \ldots, \bar{x}_{(n)}$. Todennäköisyyskuvaaja muodostuu tällöin pisteistä

$$\left\{\left(\frac{1}{n+1}, \exp\left(-e^{-\bar{x}_{(i)}}\right)\right) : i = 1, \ldots, n\right\},$$

ja kvantiilikuvaaja vastaavasti pisteistä

$$\left\{\left(-\ln\left(-\ln\frac{1}{n+1}\right), \bar{x}_{(i)}\right) : i = 1, \ldots, n\right\}.$$

Epästationaarisen GP-mallin,

$$(X_t - u(t) | X_t > u(t)) \sim G_{\boldsymbol{\theta}(t)} = G(\boldsymbol{\xi}(t), \boldsymbol{\beta}(t)),$$

kohdalla menetellään samaan tapaan. Olkoon havaittu $N = k$ mahdollisesti ajasta riippuvan tason $u(t)$ ylitystä $Y_j = X_{t_j} - u(t_j)$. Merkitään näiden havaittuja arvoja $y_j = x_{t_j} - u(t_j)$, $j = 1, \ldots, k$, missä t_j on j:nnen ylityksen sattumisaika vastaten alkuperäisen prosessin havaittua arvoa $X_{t_j} = x_{t_j}$. Koska eksponenttijakauma on yleistetyn Pareto-jauman erikoistapaus kun $\xi = 0$, GP-jakauman kohdalla on luonnollista muuntaa havainnot tähän, eli käyttää standardia eksponenttijakaumaa referenssijakaumana. Tällöin residuaalit ovat

$$\bar{y}_j = \frac{1}{\boldsymbol{\xi}(t_j)} \ln\left(1 + \boldsymbol{\xi}(t_j)\frac{y_j}{\boldsymbol{\sigma}(t_j)}\right)$$

Merkitään näiden järjestettyä otosta $\bar{y}_{(1)}, \ldots, \bar{y}_{(k)}$. Todennäköisyyskuvaaja on nyt

$$\left\{ \left(\frac{1}{k+1}, 1 - \exp\left(-\bar{y}_{(i)}\right) \right) : i = 1, \ldots, k \right\},$$

ja kvantiilikuvaaja

$$\left\{ \left(-\ln\left(1 - \frac{1}{k+1}\right), \bar{y}_{(i)} \right) : i = 1, \ldots, k \right\}.$$

2.5.5 Merenpinnan korkeuden mallintaminen epästationaarisella GEV-mallilla

Osiossa 2.2.4 todettiin, että merenpinnan korkeuden vuosimaksimien aikasarjassa (kuva 2.5) on havaittavissa mahdollinen kasvava trendi. Tarkastellaan tätä tarkemmin.

Aloitetaan yksinkertaisella lineaarisella regressiolla sovittamalla suora vuosimaksimien aikasarjaan standardilla pienimmän neliösumman (PNS) menetelmällä. Kuvaan 2.35 on piirretty havaintoihin sovitettu suora. Regressiosuoran $\boldsymbol{y} = k\boldsymbol{x} + b$ kulmakertoimen estimaatiksi saadaan $\hat{k} = 0.27$ (cm), 95 %:n luottamusvälin ollessa [0.15, 0.39]. Erityisesti luottamusväli ei sisällä nollaa, eli nollahypoteesi $H_0 : k = 0$ että datassa ei ole trendiä voidaan hylätä 95 %:n luottamustasolla. Itse asiassa data sisältää varsin vahvaa evidenssiä trendin olemassaolosta, ja nollahypoteesi voidaan hylätä korkeammillakin luottamustasoilla.

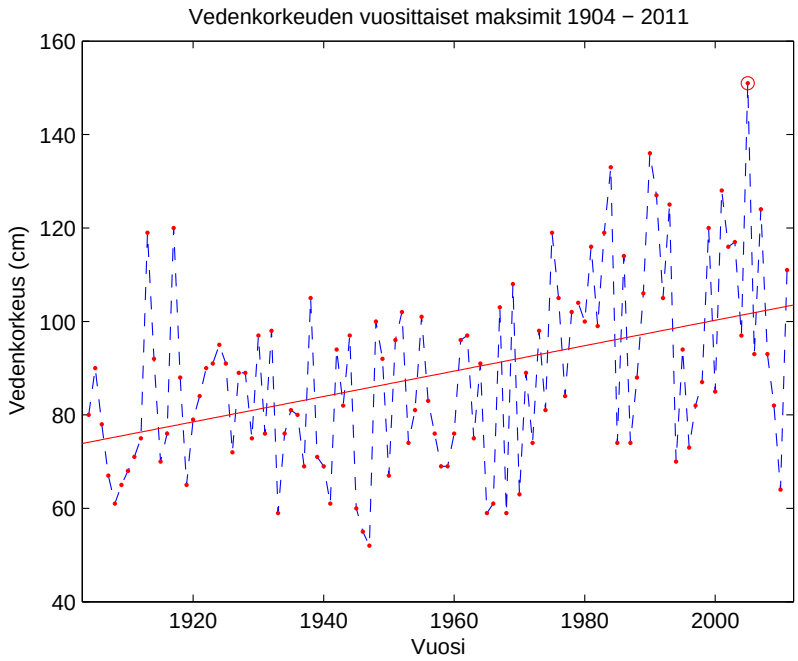
Toisaalta vuosimaksimien aikasarjaa tarkasteltaessa voidaan havaita mahdollinen muutos 1960-luvun loppupuolen tienoilla. Tätä ennen datassa ei ole havaittavissa vakuuttavaa todistusaineistoa minkäänlaisesta systemaattista trendistä. Vaihtoehtoisesti voitaisiin siis sovittaa dataan esimerkiksi malli, jossa trendi muuttuu jonakin vuosista 1965–1970. Todellakin, kun dataa edelleen tutkii lineaarisen regression keinoin sovittamalla suora toisaalta aineistoon vuosilta 1904–1969 ja toisaalta 1970–2011, havaitaan että ensimmäisessä tapauksessa suoran kulmakertoimeksi tulee lähes nolla (ei trendiä), kun taas toisessa tapauksessa kulmakerroin on selvästi positiivinen. Myös jälkimmäisessä tapauksessa tosin 95 %:n luottamusväli regressiosuoran kulmakertoimelle sisältää arvon nolla, eli trendi ei tässä tarkastelussa ole tilastollisesti merkitsevä tavanomaisella 5 %:n merkittävyystasolla – luottamusvälejä laajentamalla tai vastaavasti merkitsevyystasoa hieman kasvattamalla päätelmä kääntyisi kuitenkin päinvastaiseksi. Erityisen vahvaa tilastollista evidenssiä trendistä näyttää kuitenkin olevan vain pitkää aikaväliä tarkastellessa.

Näyttää siis siltä, että vedenkorkeuden vuosittaiset maksimiarvot olisivat keskimäärin kasvaneet pitkän, 108 vuotta kattavan tarkasteluajanjakson aikana. Sovitetaan tämän perusteella vuosimaksimidataan trendin sisältävä GEV-malli olettamalla, että vuosimaksimit X_t vuonna t noudattavat GEV-jakaumaa

$$X_t \sim H(\xi, \boldsymbol{\mu}(t), \sigma), \quad t = 1, \ldots, n,$$

missä

$$\boldsymbol{\mu}(t) = \kappa_0 + \kappa_1 t,$$



Kuva 2.35: Vedenkorkeuden vuosittaiset maksimit ja näihin sovitettu regressiosuora.

ja t valitaan vastaamaan havaintoaikasarjan ensimmäistä vuotta 1904. Parametrien SU-estimaateiksi saadaan

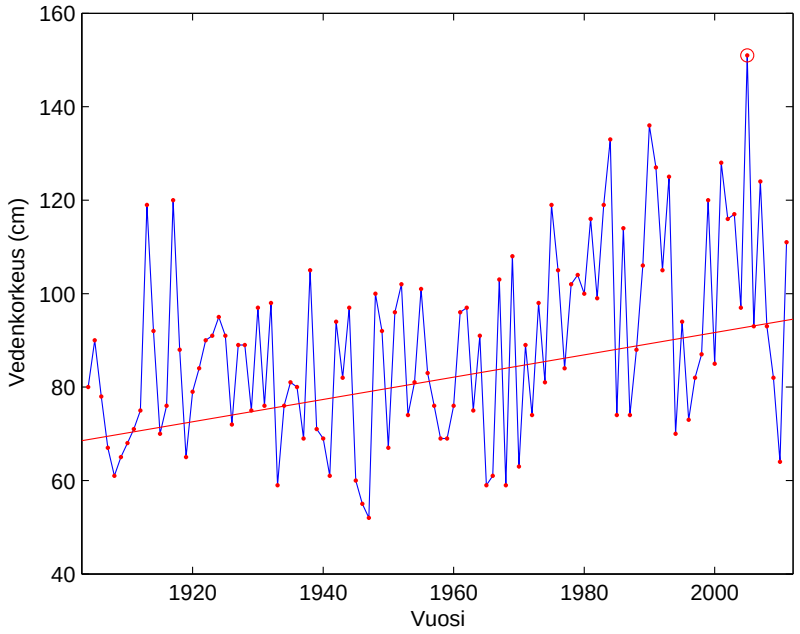
$$\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\kappa}_0, \hat{\kappa}_1, \hat{\sigma}) = (-0.195, 68.5, 0.24, 17.2)$$

ja 95 %:n approksimatiivisiksi luottamusväleiksi

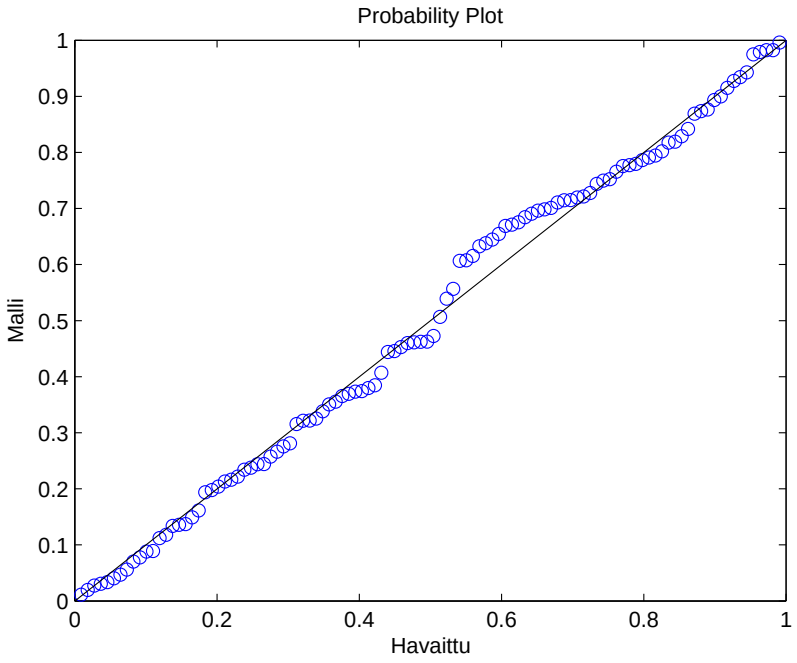
$$\hat{\boldsymbol{\theta}}_{CI} = \begin{pmatrix} \hat{\xi}^L & \hat{\kappa}_0^L & \hat{\kappa}_1^L & \hat{\sigma}^L \\ \hat{\xi}^U & \hat{\kappa}_0^U & \hat{\kappa}_1^U & \hat{\sigma}^U \end{pmatrix} = \begin{pmatrix} -0.322 & 61.5 & 0.12 & 14.8 \\ -0.068 & 75.6 & 0.35 & 20.0 \end{pmatrix}.$$

Verrattuna stationaarisen GEV-sovitteen antamiin estimaatteihin (ks. taulukko 2.3), nähdään, että trendin lokaatioparametrin suhteen sisältävä malli antaa vahvempaa evidenssiä muotoparametrin ξ negatiivisuudesta. Skaalaparametrin σ estimaatti – ja sen luottamusvälit – sen sijaan ovat hyvin samanlaisia molemmissa malleissa. Parametriestimaatin $\hat{\kappa}_1$ arvo 0.24 voidaan tulkita siten, että sovitetun mallin mukaan vuosittaisten vedenkorkeusmaksimien korkeus on kasvanut havaintojaksolla keskimäärin 0.24 cm vuodessa, eli 24 cm 100 vuodessa (vrt. yksinkertaisen lineaarisen regression estimaattiin 0.27 cm / vuosi). Kuvaan 2.36 on piirretty estimoitu suora $\hat{\boldsymbol{\mu}}(t)$ ajan funktiona; tämä on hyvin samankaltainen kuin kuvan 2.35 regressiosuora.

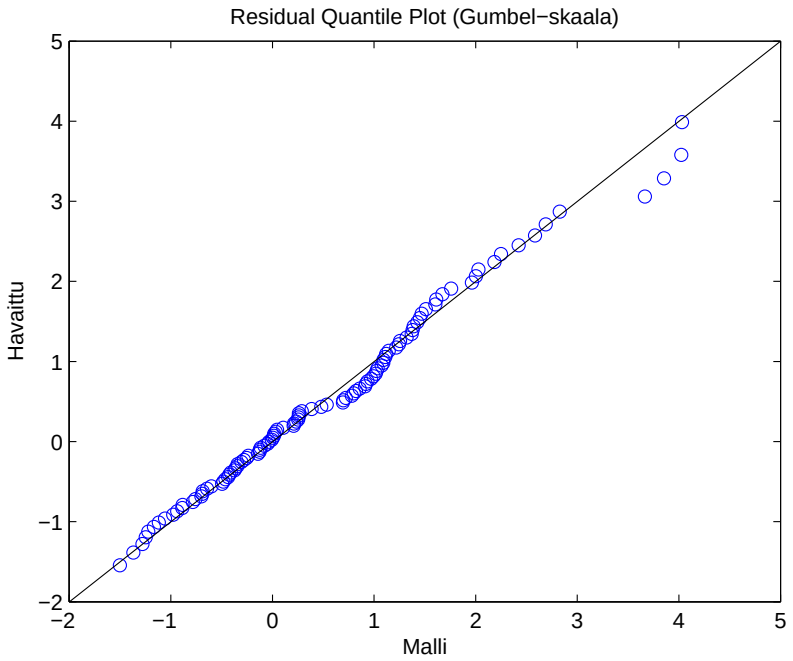
Tarkastellaan sovitetun mallin hyvyyttä todennäköisyys- ja kvantiilikuvaajien perusteella. Nämä on esitetty kuvissa 2.37 ja 2.38 . Kuvaajien perusteella ei ole syytä epäillä mallin sopivuutta havaintoihin. Sivuhuomautuksena kvantiilikuvaajasta nähdään, että tarkasteltavan mallin mukaan suurin vedenkorkeushavainto 151 cm ei ole mitenkään ”poikkeuksellinen”.



Kuva 2.36: Vedenkorkeuden vuosittaiset maksimit ja $\boldsymbol{\mu}(t)$:n estimaatti.



Kuva 2.37: Todennäköisyyskuvaaja trendin sisältävälle GEV-mallille.



Kuva 2.38: Kvantiilikuvaaja trendin sisältävälle GEV-mallille.

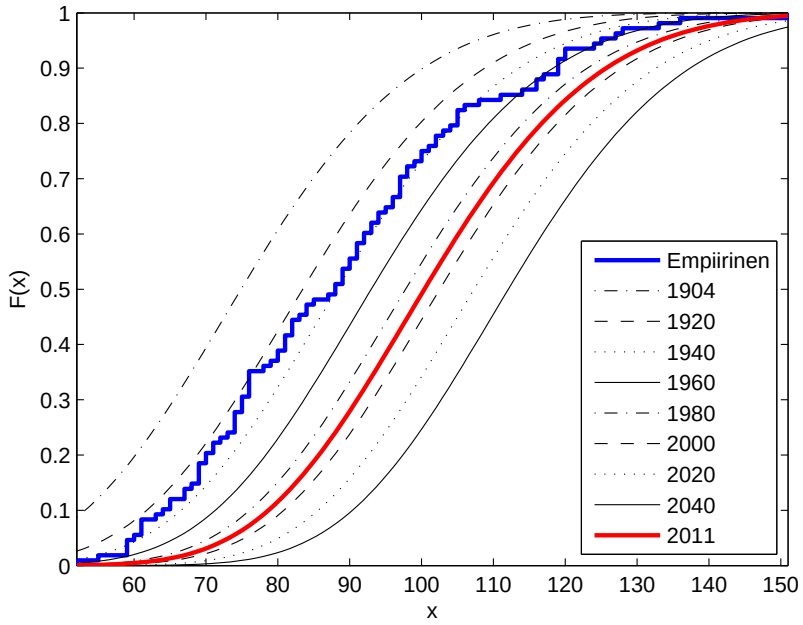
Trendin sisältävää mallia (merkitään tätä $\mathcal{M}_1$) voidaan verrata stationaariseen malliin ($\mathcal{M}_0$) suoraan uskottavuusosamäärätestin avulla, sillä malli $\mathcal{M}_1$ sisältää mallin $\mathcal{M}_0$ erikoistapauksenaan, kun $\kappa_1 = 0$. Vertailua varten muistetaan, että mallin $\mathcal{M}_0$ maksimoiduksi log-uskottavuudeksi saatiin $l_0 = -473.4$, kun mallille $\mathcal{M}_1$ vastaava log-uskottavuus on $l_1 = -466.0$. Uskottavuusosamäärätestin (ks. liite B) testisuureeksi saadaan

$$D = 2\left(l_1(\mathcal{M}_1) - l_0(\mathcal{M}_0)\right) = 2(-466.0 - (-473.3)) = 14.8,$$

kun kriittinen arvo c_α eli χ^2_1-jakauman $(1-\alpha)$-kvantiili merkitsevyystasolla $\alpha = 0.05$ on $c_{0.05} = 3.84$. Siis $D > c_{0.05}$, eli malli $\mathcal{M}_1$ selittää dataa tilastollisesti merkitsevästi paremmin 95 %:n luottamustasolla. Itse asiassa $c_{0.001} = 10.83 < D$, eli malli $\mathcal{M}_0$ hylätään mallin $\mathcal{M}_1$ eduksi myös 99.9 % luottamustasolla. Havaintoaineiston sisältämä evidenssi trendin puolesta (tai täsmällisemmin, jonkin datan ajassa muuttuvan ominaisuuden, jota lineaarinen trendi selittää) on siis erittäin vahvaa.

Mallin aikariippuvuus tarkoittaa, että vedenkorkeuden todennäköisyysjakauma muuttuu ajassa. Täten kaikki ilmiöön liittyvät todennäköisyydet, kuten ylitystodennäköisyydet ja toistumistasot, riippuvat tarkasteluhetkestä tai tarkasteluaikavälistä. Kuvaan 2.39 on piirretty mallin implikoima kertymäfunktio eri ajanhetkille, sekä havainnoista määritetty empiirinen kertymäfunktio. Empiirinen kertymäfunktio rakentuu eri vuosien havaintojen perusteella, ja kukin näistä havainnoista noudattaa mallin mukaan omaa jakaumaansa: siten empiirinen kertymäfunktio edustaa todennäköisyysjakaumien sekoitusta, eikä täsmälleen vas-

taa mitään yksittäisinä ajanhetkinä (vuosina) vallitsevista jakaumista (demografisin termein voidaan ajatella kohorttia vs. periodikohtaista jakaumaa).



Kuva 2.39: Vedenkorkeuden todennäköisyysjakauma ajan funktiona trendin sisältävän GEV-mallin mukaan.

Taulukossa 2.10 on esitetty suurimman toistaiseksi mitatun merenpinnan tason $x_p = 151$ cm ylittämisen (vuosittaisia) todennäköisyyksiä p_t muutamille vuosille t, sekä vertailuna osiossa 2.2.4 stationaarisella GEV-mallilla saatu (ajasta riippumaton) todennäköisyys, $p(\mathcal{M}_0)$.

Taulukko 2.10: Suurimman havaitun vedenkorkeustason ylitystodennäköisyyksiä p joillekin vuosille trendin sisältävässä GEV-mallissa.

x_p	$p(\mathcal{M}_0)$	p_{1904}	p_{1960}	p_{2011}	p_{2020}
151 (cm)	0.0063	$1.3 \cdot 10^{-6}$	$4.4 \cdot 10^{-4}$	0.0052	0.0073

2.6 Pisteprosessit

Kappaleessa 1.6 nähtiin, että pisteprosessimalli korkean tason ylitteille tarjoaa ääriarvoanalyysille yhtenäisen viitekehyksen: iid datalle formuloitu POT-malli yhdistää yhteen malliin sekä GEV-jakaumaan perustuvan mallin blokkimaksimeille että GP-jakaumaan perustuvan mallin ylitteiden suuruuksille.

Osiossa 1.6.2 käsitellyn POT-mallin taustalla oleva (homogeeninen) Poisson-prosessi ylityksien sattumiselle saadaan asymptoottisena rajatuloksena kun tar-

kastellaan iid datan ylitteitä, tai myös stationaarisen sarjan tapauksessa kun tarkastellaan klusterimaksimeita; ks. [4], [23]. Tämä motivoi tuttuun tapaan käyttämään tulosta approksimaationa korkealle mutta äärelliselle kynnystasolle u, perustellen mallin sovittamisen dataan.

POT-mallia voidaan ajatella useasta lähtökohdasta käsin: Yksi tapa on ajatella korkean tason u ylityksien tapahtuvan homogeenisen Poisson-prosessin mukaisesti intensiteetillä λ, ja ylitteiden suuruuksien olevan yleistettyä Pareto-jakaumaa noudattavia, ylitysajoista riippumattomia iid satunnaismuuttujia. Tätä kutsutaan joskus Poisson-GP–malliksi. Toinen tapa on käsitellä ylitteitä pisteinä kaksiulotteisessa avaruudessa ja mallintaa niitä suoraan 2-ulotteisella epähomogeenisellä Poisson-prosessilla.

Poisson-GP –lähestymistapa mahdollistaa Poisson- ja GP-komponenttien estimoimisen datasta erikseen ("ortogonaalinen" lähestymistapa). Tämä saattaa joskus olla estimoinnin kannalta mukavaa. Pisteprosessilähestymistapa vaatii kaksiulotteisen epähomogeenisen Poisson-prosessin sovittamista dataan. Iid datan tapauksessa molemmat tuottavat teoriassa saman lopputuloksen. Pisteprosessilähestymistavan ehdottomana etuna on kuitenkin sen joustavuus ja helppo yleistäminen sisältämään aikariippuvan intensiteetin ja ulkoisia selittäviä muuttujia. Ortogonaalisessa lähestymistavassa tämä ei yleensä onnistu luonnollisella tavalla, ja on usein hankalaa. Poisson-pisteprosessissa parametreilla ξ, μ, σ ei ole teoreettista riippuvuutta valitusta kynnystasosta u, kun GP-jakauman skaalaparametri β puolestaan on suoraan u:n funktio ja siten muuttuu tason mukana.

Perus-POT-mallin Poisson-pisteprosessi on homogeeninen ajan suhteen, mutta epähomogeeninen ylitteiden suuruuden (vertikaalisen dimension) suhteen, koska korkeammat ylitteet ovat epätodennäköisempiä. Kun POT-mallin (intensiteetin) parametrit asetetaan riippumaan ajasta tai selittävistä muuttujista, saadaan tarkasteltavan prosessin epästationaarisuus luonnollisella tavalla huomioitua. Tällaiset pisteprosessit ovat epähomogeenisiä myös ajan suhteen.

2.6.1 Suurimman uskottavuuden menetelmä pisteprosesseille

Kiinnitetään tila-avaruus $E = \mathbb{R}_+ \times S$ vastaavalla Borel-σ-algebralla $\mathscr{E}$, ja tarkastellaan pisteprosessin N realisaatioita tila-avaruuden E osajoukossa $\mathcal{R} \subset E$ (sigma-algebralla $\mathscr{E}(\mathcal{R}) = \sigma(\mathcal{R})$). Realisaatiot koostuvat pisteiden lukumäärästä ja niiden paikoista. Ylitysten pisteprosessin tapauksessa pisteet ovat siis valitun tason u ylittäviä pisteitä $(T_j, \tilde{X}_j) \in \mathcal{R}$ alla olevasta satunnaismuuttujajonosta $(X_1, \ldots, X_n)$, ja realisaatiot ovat ylitteiden joukkoja $\{(T_j, \tilde{X}_j) : j = 1, \ldots, K_u\} \subset \mathscr{E}(\mathcal{R})$, missä K_u on ylityksien lukumäärä. T_j tulkitaan j:nnen ylityksen sattumisajaksi ja $\tilde{X}_j$ vastaavasti ylitteen suuruudeksi.

Esitetään seuraavassa hahmotelma uskottavuusfunktion johtamisesta Poisson-pisteprosessille N. Pisteprosessien uskottavuusfunktioista tarkemmin ks. [30, luvut 7.1–7.3]. Merkitään yksittäistä pistettä $\xi_j = (T_j, \tilde{X}_j)$ ja koko realisaatiota $\Xi = (\xi_1, \ldots, \xi_{K_u})$. Pistejoukon Ξ todennäköisyys voidaan kirjoittaa

$$\mathbb{P}(N = \Xi) = \mathbb{P}(K_u = k)\mathbb{P}(N = (\xi_1, \ldots, \xi_k) | K_u = k). \tag{2.19}$$

Ensimmäinen todennäköisyys oikealla on nyt

$$\mathbb{P}(K_u = k) = p_{K_u}(k) = e^{-\Lambda(\mathcal{R})} \frac{(\Lambda(\mathcal{R}))^k}{k!}, \tag{2.20}$$

koska N on Poisson-pisteprosessi. Geneeriselle satunnaismuuttujalle $\xi \sim \xi_j$ pätee

$$\mathbb{P}(\xi \in A) = \frac{\Lambda(A)}{\Lambda(\mathcal{R})}, \quad A \in \mathscr{E}(\mathcal{R}),$$

kun $\Lambda(\mathcal{R}) < \infty$ (vrt. liitteen D.1 konstruktio). Intensiteettimitta on nyt esitettävissä muodossa $\Lambda(A) = \int_A \lambda(x)\,\mathrm{d}x$, jolloin derivaattana Lebesgue-mitan suhteen saadaan ξ:n tiheysfunktioksi

$$p_\xi(x) = \frac{d\mathbb{P}(\xi \in x)}{dx} = \frac{1}{\Lambda(\mathcal{R})} \frac{d\Lambda(x)}{dx} = \frac{\lambda(x)}{\int_{\mathcal{R}} \lambda(s)\,\mathrm{d}s}, \quad x \in \mathscr{E}(\mathcal{R}).$$

Satunnaismuuttujien ξ ehdollisesta riippumattomuudesta seuraa, että N:n ehdollinen tiheysfunktio on muotoa

$$p_{N|K_u}((\xi_1, \ldots, \xi_k)|k) = \prod_{j=1}^{k} p_\xi(\xi_j), \tag{2.21}$$

jolloin, sijoittamalla yhtälöt (2.20) ja (2.21) yhtälöön (2.19), saadaan N:n tiheysfunktioksi realisaatiossa $N = \Xi = (\xi_1, \ldots, \xi_k)$

$$\begin{aligned} p_N(\Xi) &= p_{K_u}(k) p_{N|K_u}\left((\xi_1, \ldots, \xi_n)|k\right) \\ &= \exp\left\{-\Lambda\left(\mathcal{R}\right)\right\} \frac{(\Lambda(\mathcal{R}))^k}{k!} \prod_{i=1}^{k} \frac{\lambda(\xi_j)}{\Lambda(\mathcal{R})} \\ &= \frac{1}{k!} \exp\left\{-\Lambda\left(\mathcal{R}\right)\right\} \prod_{j=1}^{k} \lambda(\xi_j), \end{aligned} \tag{2.22}$$

kun $k \geq 0$ (jos $K_u = k = 0$, niin tulo yllä tulkitaan tavalliseen tapaan 1:ksi). Uskottavuusfunktio on nyt tiheysfunktio (2.22) evaluoituna havaitulla otoksella ja tulkittuna (intensiteetin) parametrien funktioksi; ks. seuraavat alaosiot. Tekijä $\frac{1}{k!}$ tiheysfunktiossa on vakio kiinnitetyllä otoksella, eikä vaikuta parametrien estimointiin. Täten se jätetään jatkossa pois uskottavuusfunktiosta.

POT-malli. Olkoon ylitteiden havaittu määrä otoksessa $K_u = k$, havaitut ylitysajat $t_1, \ldots, t_n$ ja ylitteiden suuruudet $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \ldots, \tilde{x}_k)$, kun taso u on kiinnitetty. Log-uskottavuus POT-mallin Poisson-pisteprosessin tapauksessa voidaan kirjoittaa

$$\begin{aligned} L(\boldsymbol{\theta}; \tilde{\boldsymbol{x}}) &= \exp\left\{-\Lambda\left((0, n] \times (u, \infty)\right)\right\} \prod_{i=1}^{k} \lambda(\tilde{x}_i) \\ &= \exp\left\{-n\tau(u)\right\} \prod_{i=1}^{k} \lambda(\tilde{x}_i), \end{aligned}$$

missä $\boldsymbol{\theta} = (\xi, \mu, \sigma)$; notaation osalta ks. osio 1.6.2. Log-uskottavuus on vastaavasti

$$\begin{aligned} l(\boldsymbol{\theta}; \tilde{\boldsymbol{x}}) &= -n\tau(u) + \sum_{i=1}^{k} \ln \lambda(\tilde{x}_i) = n \ln H_{\xi,\sigma,\mu}(u) + \sum_{i=1}^{k} \ln \lambda(\tilde{x}_i) \\ &= -n \left(1 + \xi \frac{u-\mu}{\sigma}\right)^{-1/\xi} + \sum_{i=1}^{k} \ln \left\{ \frac{1}{\sigma} \left(1 + \xi \frac{\tilde{x}_i - \mu}{\sigma}\right)^{-1/\xi - 1} \right\}, \end{aligned} \tag{2.23}$$

kun $1 + \xi(\tilde{x}_i - \mu)/\sigma > 0$ kaikilla $i = 1, \ldots, k$ (ja luonnollisesti myös $1 + \xi(u - \mu)/\sigma > 0$, $\sigma > 0$). Parametrit saadaan estimoitua maksimoimalla (log-)uskottavuusfunktio numeerisesti.

Aikadimension suhteen epähomogeeniset mallit. Tarkastellaan mallia, jossa Poisson-pisteprosessin intensiteetti $\lambda(t, x)$ riippuu sekä t:stä että x:stä. Edellistä kohtaa mukaillen uskottavuusfunktio on nyt

$$\begin{aligned} L(\boldsymbol{\kappa}; \tilde{\boldsymbol{x}}) &= \exp\left\{-\Lambda\left((0, n] \times (u, \infty)\right)\right\} \prod_{i=1}^{k} \lambda(t_i, \tilde{x}_i) \\ &= \exp\left\{ -\int_0^n \int_u^\infty \lambda(t, y) \, \mathrm{d}y \, \mathrm{d}t \right\} \prod_{i=1}^{k} \lambda(t_i, \tilde{x}_i), \end{aligned}$$

missä parametrivektoriin $\boldsymbol{\kappa}$ on kerätty intensiteetin parametreille $\boldsymbol{\theta}(t) = (\boldsymbol{\xi}(t), \boldsymbol{\mu}(t), \boldsymbol{\sigma}(t))$ käytettyjen mallien parametrit (ks. osio 2.5). Log-uskottavuus on

$$\begin{aligned} l(\boldsymbol{\kappa}; \tilde{\boldsymbol{x}}) &= -\int_0^n \int_u^\infty \lambda(t, y) \, \mathrm{d}y \, \mathrm{d}t + \sum_{i=1}^{k} \lambda(t_i, \tilde{x}_i), \\ &= -\int_0^n \left(1 + \boldsymbol{\xi}(t) \frac{u - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)}\right)^{-1/\boldsymbol{\xi}(t)} \mathrm{d}t \\ &\quad + \sum_{i=1}^{k} \ln \left\{ \frac{1}{\boldsymbol{\sigma}(t_i)} \left(1 + \boldsymbol{\xi}(t_i) \frac{\tilde{x}_i - \boldsymbol{\mu}(t_i)}{\boldsymbol{\sigma}(t_i)}\right)^{-1/\boldsymbol{\xi}(t_i) - 1} \right\}, \end{aligned} \tag{2.24}$$

kun $1 + \boldsymbol{\xi}(t)(\tilde{x}_i - \boldsymbol{\mu}(t))/\boldsymbol{\sigma}(t) > 0$ kaikilla $(t, x) = (t_i, \tilde{x}_i)$, $i = 1, \ldots, k$.

2.6.2 Mallidiagnostiikkaa

2.6.2.1 Ylitteiden suuruudet

Todennäköisyys- ja kvantiilikuvaajien muodostamiseksi pisteprosessimallille muistetaan ensinnäkin osiosta 1.6.3, että ylitteet noudattavat GP-jakaumaa:

$$\begin{aligned} \bar{F}_{u(t)}(x) &= \mathbb{P}(X_t > u(t) + x | X_t > u(t)) \\ &= \left(1 + \boldsymbol{\xi}(t) \frac{x}{\boldsymbol{\sigma}(t) + \boldsymbol{\xi}(t)(u(t) - \boldsymbol{\mu}(t))}\right)^{-1/\boldsymbol{\xi}} \\ &= \bar{F}_{\boldsymbol{\xi}(t), \boldsymbol{\beta}(t)}(x), \end{aligned}$$

missä $\boldsymbol{\beta}(t) = \boldsymbol{\sigma}(t)+\boldsymbol{\xi}(t)(u(t)-\boldsymbol{\mu}(t))$. Nyt voidaan suoraan hyödyntää epästationaarista GP-jakaumaa koskevia tuloksia osiosta 2.5.4. Muunnetaan ylitteet $Y_j = X_{T_j} - u(T_j) = \tilde{X}_j - u(T_j)$ seuraavasti:

$$\bar{Y}_j = \frac{1}{\boldsymbol{\xi}(T_j)} \ln\left(1 + \boldsymbol{\xi}(T_j)\frac{Y_j}{\boldsymbol{\sigma}(T_j) + \boldsymbol{\xi}(T_j)(u(T_j) - \boldsymbol{\mu}(T_j))}\right).$$

Nämä noudattavat nyt standardia eksponenttijakaumaa. Olkoon havaittu ylitteet $\boldsymbol{y} = (y_1, \ldots, y_k)$ ajanhetkillä $t_1, \ldots, t_k$, ja merkitään näitä vastaavien residuaalien $\bar{\boldsymbol{y}} = (\bar{y}_1, \ldots, \bar{y}_k)$ järjestettyä otosta $\bar{y}_{(1)}, \ldots, \bar{y}_{(k)}$. Kuten osiossa 2.5.4, todennäköisyyskuvaaja on tällöin

$$\left\{\left(\frac{1}{k+1}, 1 - \exp\left(-\bar{y}_{(i)}\right)\right) : i = 1, \ldots, k\right\},$$

ja kvantiilikuvaaja

$$\left\{\left(-\ln\left(1 - \frac{1}{k+1}\right), \bar{y}_{(i)}\right) : i = 1, \ldots, k\right\}.$$

2.6.2.2 Ylitteiden sattumisajat

Edellisen osion tarkastelu koski ylitteiden ehdollista suuruusjakaumaa (ylitejakaumaa) tarkasteltuna erikseen ylitteiden sattumisprosessista, eli ilman aikaulottuvuutta. Tarkastellaan seuraavaksi ylitteiden sattumisaikoja, ja perus-Poisson-pisteprosessimallin taustalla olevaa oletusta ylitteiden lukumäärien Poisson-jakautuneisuudesta. Seuraava perustuu lähteeseen [31]; ks. myös [30, luku 7.4].

Oletetaan, että kynnystason u ylityksien sattumisajat $\{t_i\}$ on generoinut (yksiulotteinen) prosessi N intensiteetillä $\lambda(t, u) = \lambda_u(t)$, ja tarkastellaan integraalia

$$\Lambda_u(t) = \int_0^t \lambda_u(s)\,\mathrm{d}s.$$

Tämä on kasvava, koska $\lambda(t)$ on ei-negatiivinen. Määritellään stokastinen aikamuunnos $t \mapsto \tau$, missä $\tau = \Lambda_u(t)$; joukko $\{t_i : i = 1, \ldots, n\}$ kuvautuu nyt joukoksi $\{\tau_i : i = 1, \ldots, n\}$. Tiedetään, että muunnettu prosessi $\tilde{N}(t) := N(\tau) = N(\Lambda_u^{-1}(t))$ on Poisson-prosessi vakiointensiteetillä 1 (ks. liite C). Toisin sanoen, muunnetut sattumisajat $\{\tau_i = \Lambda_u(t_i)\}$ ovat realisaatio Poisson-prosessista intensiteetillä 1, jos (ja vain jos) alkuperäiset ajat $\{t_i\}$ ovat realisaatio prosessista intensiteetillä $\Lambda_u(t)$. Jos siis havaitun datan perusteella estimoitu ylitysintensiteetti $\hat{\lambda}_u(t)$ vastaa hyvin todellista sattumisintensiteettiä $\lambda_u(t)$, tulisi muunnosten (τ_i) käyttäytyä kuin homogeeninen Poisson-prosessi.

Kenties vielä havainnollisempaa on tarkastella odotusaikoja: Poisson-jakauman perusominaisuuksien perusteella tiedetään, että mikäli sattumisajat noudattavat Poisson-prosessia, ovat peräkkäisten tapahtumien väliset ajat (inter-arrival times) eli odotusajat eksponenttijakautuneita satunnaismuuttujia. Tarkastellaan muunnettuja odotusaikoja

$$Y_k = \tau_k - \tau_{k-1} = \Lambda_u(t_k) - \Lambda_u(t_{k-1}) = \int_{t_{k-1}}^{t_k} \lambda_u(s)\,\mathrm{d}s, \quad k = 1, \ldots, n.$$

Mikäli (τ_k) ovat Poisson-jakautuneita, ovat (Y_k) iid eksponenttijakautuneita satunnaismuuttujia parametrilla 1. Edelleen suureet

$$U_k = 1 - \exp(-Y_k) \tag{2.25}$$

ovat tällöin riippumattomia ja tasajakautuneita välille [0,1). Kolmogorov-Smirnov-testiä voidaan nyt käyttää testaamaan otoksen (U_k) tasajakautuneisuutta vertaamalla otoksen empiiristä kertymäfunktiota referenssijakauman eli $T(0,1)$-tasajakauman kertymäfunktiota vasten.

Odotusaikojen riippumattomuutta voidaan testata myös piirtämällä muunnetut odotusajat U_{k+1} vierekkäisiä odotusaikoja U_k vasten. Mikäli (vierekkäiset) odotusajat ovat riippumattomia, tulisi pisteiden $\{(U_k, U_{k+1}) : k = 1, \ldots, n\}$ olla tasan jakautuneita alueessa $[0,1)\times[0,1)$, eli piirretyjen pisteiden tulisi jakaantua täysin satunnaisesti yksikköneliöön.

2.6.3 Merenpinnan korkeuden mallintaminen pisteprosesseja käyttäen

Kuten aiemmin on mainittu, pisteprosessilähestymistapa tarjoaa GEV- tai GPD-malleja luonnollisemman tavan yleistää mallien parametrit ajasta tai muista muuttujista riippuviksi ja tarkastella regressiomallien upottamista tilastolliseen ääriarvoanalyysiin.

Keskitytään tutkimaan erilaisia mallirakenteita vedenkorkeuden ääritasoille ja tarkastelemaan näiden (keskinäistä) hyvyyttä.

Aloitetaan tarkastelemalla perusmuotoista POT-mallia, joka perustuu iid-oletukseen ja jossa (korkean tason u) ylityksien sattuminen noudattaa siis homogeenista Poisson-prosessia, ja ylitteet (t,x)-tasossa kaksiulotteista, ajan suhteen homogeenista Poisson-pisteprosessia. Poisson-pisteprosessin intensiteetti pisteessä (t,x) on

$$\lambda(t,x) \equiv \lambda(x) = \frac{1}{\sigma}\left(1 + \xi\frac{x-\mu}{\sigma}\right)^{-1/\xi-1}.$$

Otetaan kynnystasoiksi osiossa 2.3.5 saadut $u = 96$ cm ja $u = 108$ cm ja tarkastellaan näiden tasojen ylitteitä päivätason datassa. Maksimoimalla POT-mallin log-uskottavuusfunktio (2.23) parametrien SU-estimaateiksi tason $u = 96$ tapauksessa saadaan

$$\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\mu}, \hat{\sigma}) = (-0.176, 94.1, 14.9),$$

ja maksimoiduksi log-uskottavuudeksi tulee -440.2. Tasolle $u = 108$ vastaavasti

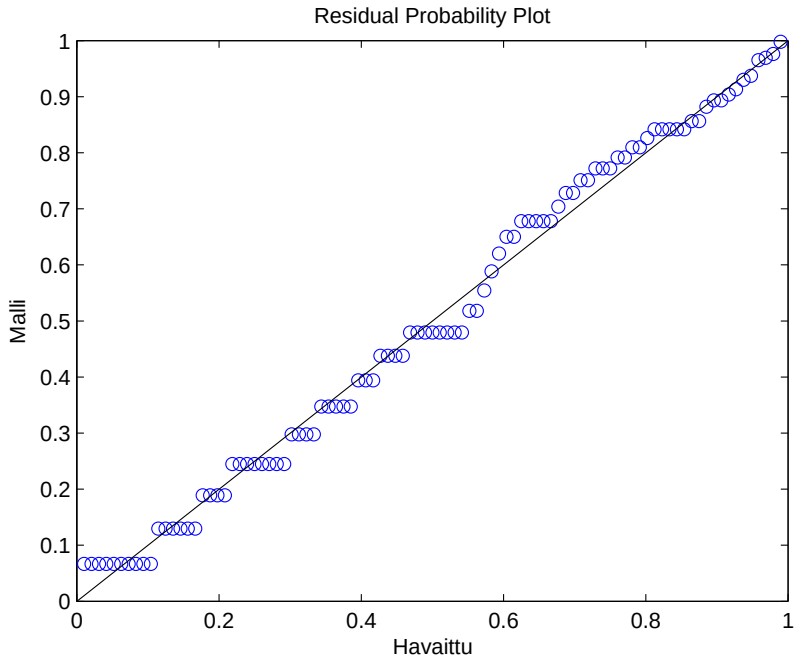
$$\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\mu}, \hat{\sigma}) = (-0.208, 93.1, 16.3),$$

ja log-uskottavuus on -210.1 (log-uskottavuuksia ei tietenkään voi verrata keskenään estimoinnissa käytetyn datan erilaisuudesta johtuen).

Huomataan, että muotoparametrien ξ estimaattien arvot ovat täsmälleen samat kuin osiossa 2.3.5 saadut ylitemenetelmän vastaavat, kuten teorian perusteella

odotettiinkin. Lisäksi GP-jakauman skaalaparametrin ja POT-mallin parametrien välillä on yhteys $\beta = \sigma + \xi(u - \mu)$; tämä antaa $\hat{\beta}$:n arvoiksi 14.60 ja 13.16, kuten pitääkin.

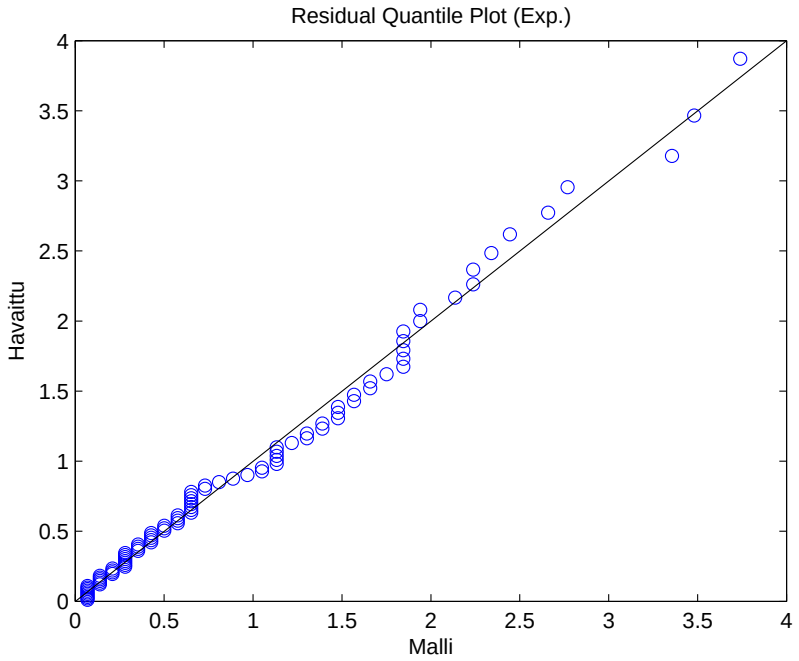
Kuvissa 2.40 ja 2.41 on todennäköisyys- ja kvantiilikuvaaja tason $u = 96$ ylitteisiin sovitetulle mallille, ja kuvissa 2.42 ja 2.43 vastaavat korkeamman tason $u = 108$ ylitteisiin perustuvalle mallille. Kuvaajien perusteella malli näyttää sopivan havaintoihin kohtuullisen hyvin.



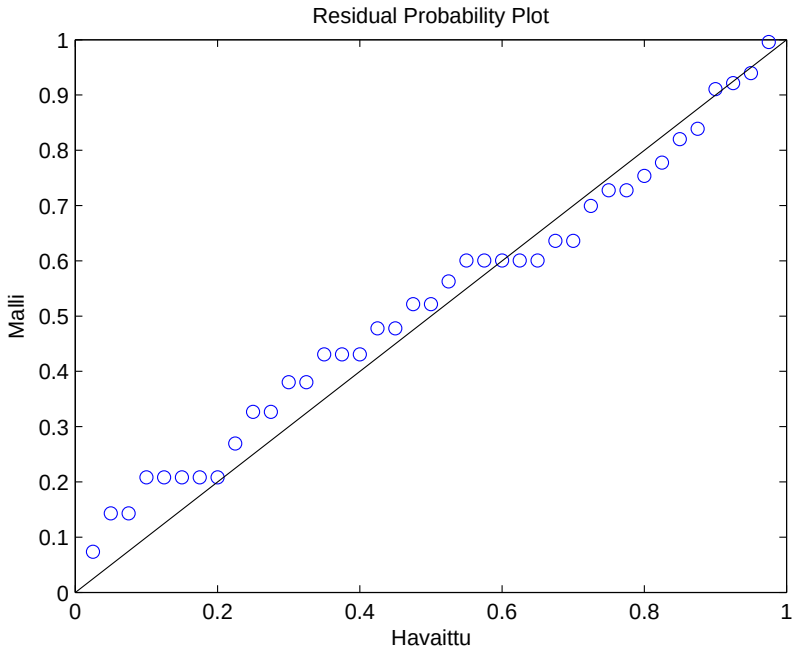
Kuva 2.40: Todennäköisyyskuvaaja POT-mallille ($u = 96$).

Ylitemenetelmän yhteydestä osiosta 2.3.5 muistetaan kuitenkin, etteivät korkean tason ylittävien vedenkorkeushavaintojen sattumisajat näyttäneet noudattavan homogeenista Poisson-prosessia, mikä on implisiittisesti GP-jakaumaan perustuvan ylitemenetelmän ja eksplisiittisesti perus-Pot-mallin oletuksena. Tämä sotii ylityksien iid-oletusta vastaan.

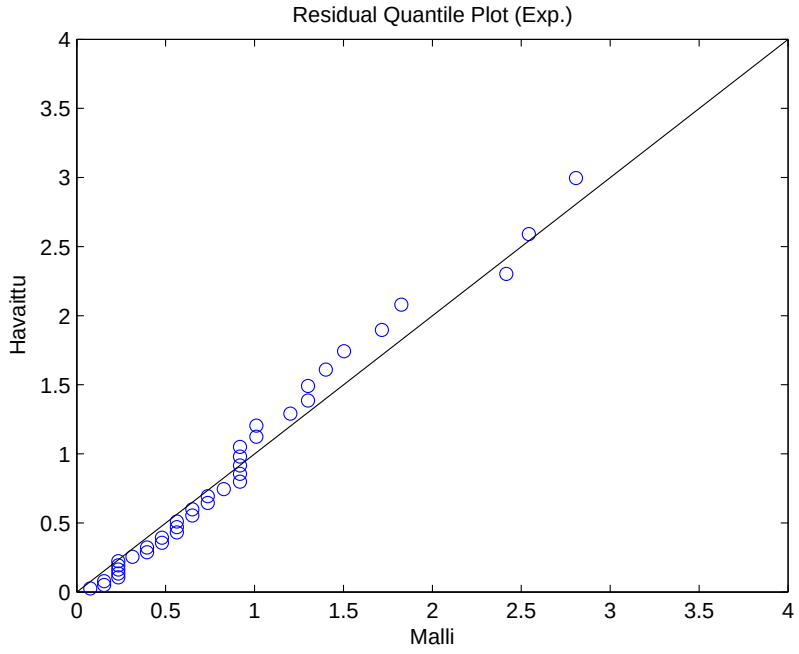
Poisson-jakauman perusominaisuuksien perusteella tiedetään, että mikäli ylitykset ovat Poisson-jakautuneita, ovat peräkkäisten ylitysaikojen välit eli odotusajat eksponenttijakautuneita satunnaismuuttujia. Poisson-oletuksen testaaminen voidaan perustaa tähän edellä alaosiossa 2.6.2.2 kuvatulla tavalla. Kuviin 2.44 ja 2.45 on piirretty muunnettuun odotusaikaan perustuvan suureen U_k empiirinen jakauma. Mikäli Poisson-oletus pätee, tulisi muunnettujen intervallien U_k olla tasajakautuneita välille $[0, 1)$, eli kuvaajan pisteiden tulisi osua lähelle yksikködiagonaalia (punaiset katkoviivat kuvissa osoittavat 95 %:n luottamusvälit). Nähdään, että tason $u = 96$ kohdalla oletus voidaan hylätä. Tason $u = 108$ kohdalla ollaan hieman lähempänä tasajakautuneisuutta, mutta johtopäätös pysyy samana.



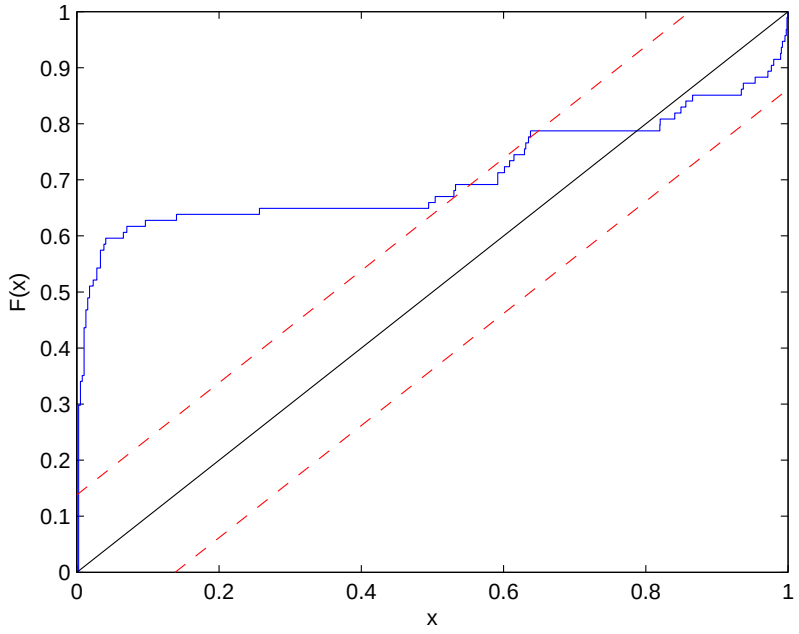
Kuva 2.41: Kvantiilikuvaaja POT-mallille ($u = 96$).



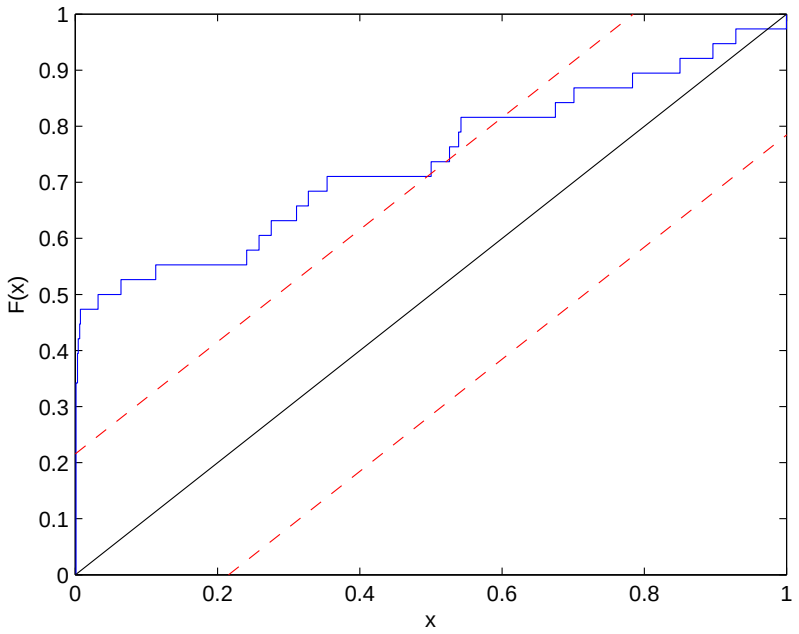
Kuva 2.42: Todennäköisyyskuvaaja POT-mallille ($u = 108$).



Kuva 2.43: Kvantiilikuvaaja POT-mallille ($u = 108$).



Kuva 2.44: Muunnettuihin odotusaikoihin perustuvan suureen U_k empiirinen jakauma (sininen viiva) tasajakaumaa vastaan luottamusväleineen; $u = 96$.



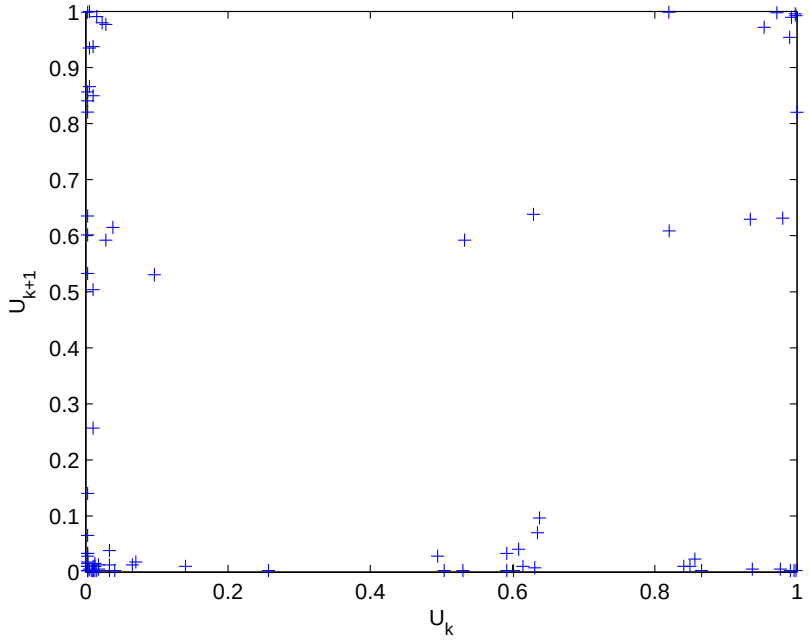
Kuva 2.45: Muunnettuihin odotusaikoihin perustuvan suureen U_k empiirinen jakauma (sininen viiva) tasajakaumaa vastaan luottamusväleineen; $u = 108$.

Piirtämällä suureet U_{k+1} vierekkäisiä suureita U_k vasten voidaan testata odotusaikojen riippumattomuutta. Mikäli viereiset odotusajat ovat riippumattomia, tulisi pisteiden (U_k, U_{k+1}) jakaantua satunnaisesti alueeseen $[0, 1) \times [0, 1)$ eli olla tasajakautuneita kuvan muodostamassa yksikköneliössä. Kuviin 2.46 ja 2.47 on piirretty pisteet tasoja $u = 96$ ja $u = 108$ vastaten. Yksikköneliön reunamille kasaantuvien pisteiden lukumäärä merkille pannen nähdään, että pisteiden jakauma ei vaikuta kovin satunnaiselta – varsinkaan ensimmäisessä kuvassa – mikä vahvistaa johtopäätöstä, etteivät vierekkäiset ylitykset ole riippumattomia.

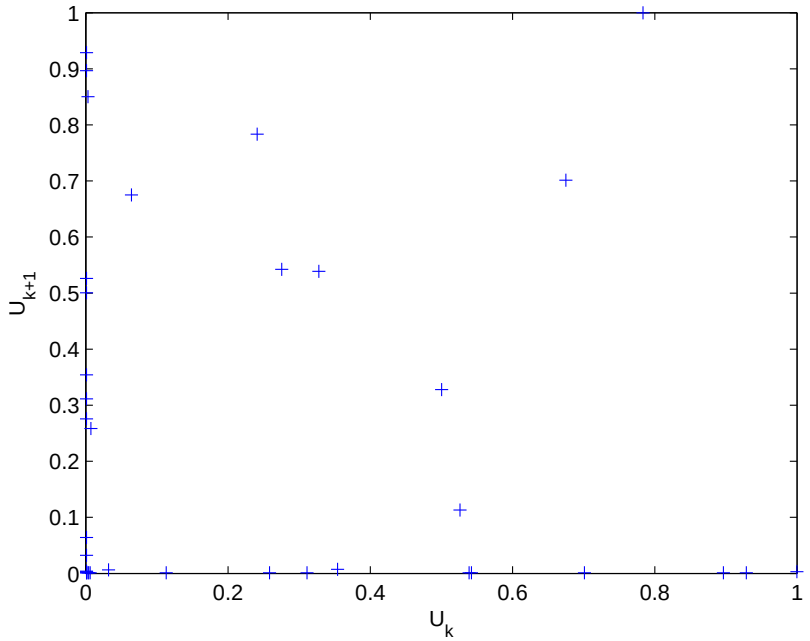
Ylitteiden riippumattomuuden (ja Poisson-oletuksen) ilmeisestä toteutumattomuudesta huolimatta sovitettu ajan suhteen homogeeniseen Poisson-pisteprosessiin perustuva malli näytti sopivan havaintoihin varsin hyvin. Luovutaan seuraavassa prosessin homogeenisuusoletuksesta aikadimension suhteen, ja tarkastellaan malleja joissa ylitykset sattuvat epähomogeenisen Poisson-pisteprosessin mukaisesti.

2.6.3.1 Aikariippuva intensiteetti

Kuvataan ylitteitä Poisson-pisteprosessina tila-avaruudessa $E = (0, n] \times (u, \infty)$ intensiteettimitalla $\Lambda(A) = \int_A \lambda(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$. Siis muotoa $A = (t_1, t_2) \times (u, \infty) \subset E$



Kuva 2.46: U_k vs. U_{k+1}; $u = 96$.



Kuva 2.47: U_k vs. U_{k+1}; $u = 108$.

olevilla joukoilla

$$\Lambda(A) = \int_{t_1}^{t_2} \int_{u}^{\infty} \lambda(t, y)\, \mathrm{d}y\, \mathrm{d}t,$$

missä intensiteetti pisteessä (t, x) on

$$\lambda(t, x) = \frac{1}{\boldsymbol{\sigma}(t)} \left(1 + \boldsymbol{\xi}(t) \frac{x - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)}\right)^{-1/\boldsymbol{\xi}(t) - 1}.$$

Tutkitaan seuraavia malliparametrisointeja:

- malli $\mathcal{M}_1$: $\boldsymbol{\theta} = (\xi, \boldsymbol{\mu}(t), \sigma)$, missä $\boldsymbol{\mu}(t) = \kappa_0 + \kappa_1 t$,
- malli $\mathcal{M}_2$: $\boldsymbol{\theta} = (\xi, \mu, \boldsymbol{\sigma}(t))$, missä $\boldsymbol{\sigma}(t) = e^{\kappa_0 + \kappa_1 t}$.

Vertailukohtana käytetään edellisen osion ajan suhteen homogeenista POT-mallia, jota merkitään $\mathcal{M}_0$. Tarkastellaan jatkossa tilan säästämiseksi vain tasoa $u = 96$ cm ja sen ylittävien havaintojen muodostamaa dataa.

Malli $\boldsymbol{\mathcal{M}_1}$. Maksimoimalla log-uskottavuus saadaan malliparametrien SU-estimaateiksi

$$\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\kappa}_0, \hat{\kappa}_1, \hat{\sigma}) = (-0.180, 68.7, 0.39, 13.9).$$

Maksimoidun log-uskottavuuden arvo on -406.6.

Verrataan mallia $\mathcal{M}_1$ malliin $\mathcal{M}_0$, joka on siis edellisen erikoistapaus. Uskottavuusosamäärätestin testisuureen arvoksi saadaan

$$D = 2\left(l_1(\mathcal{M}_1) - l_0(\mathcal{M}_0)\right) = 2(-406.6 - (-440.2)) = 67.2,$$

mikä on erittäin suuri χ^2_1-jakauman skaalalla. Esimerkiksi luottamustasoa 99.9 % vastaava kriittinen taso on $c_{0.001} = 10.83 < D$; trendin sisältävä malli $\mathcal{M}_1$ on siis merkitsevästi mallia $\mathcal{M}_0$ parempi (havaitun datan selittämisen mielessä) käytännössä millä tahansa luottamustasolla.

Alla kuvissa 2.48 ja 2.49 on esitetty todennäköisyyskuvaaja ja kvantiilikuvaaja estimoidulle mallille. Näiden perusteella malli sopii havaintoihin hyvin.

Malli $\boldsymbol{\mathcal{M}_2}$. Suurimman uskottavuuden menetelmällä kakkosmallin paremetrien estimaateiksi saadaan

$$\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\mu}, \hat{\kappa}_0, \hat{\kappa}_1) = (-0.172, 92.1, 2.0, 0.0092),$$

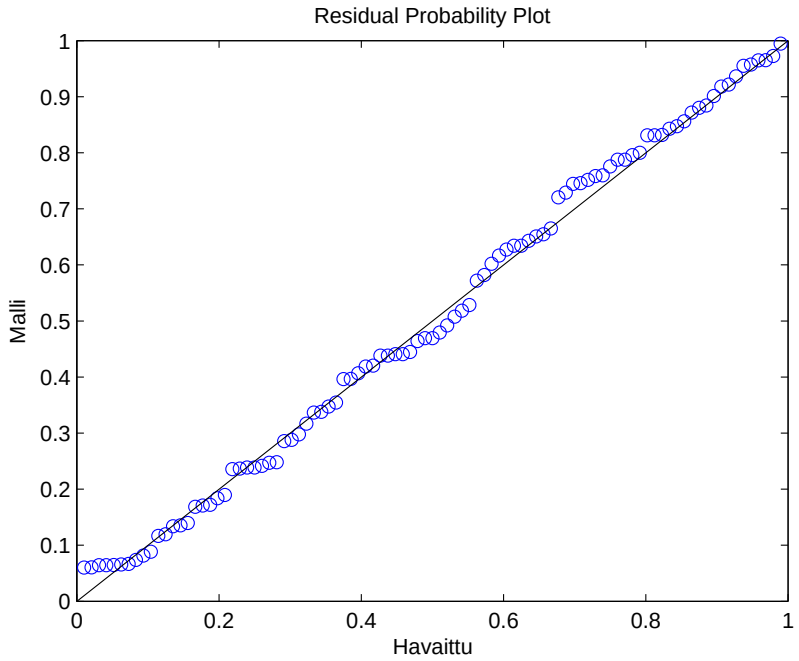
ja maksimaalinen log-uskottavuus on -435.6.

Perusmalliin $\mathcal{M}_0$ verrattaessa testisuureeksi saadaan

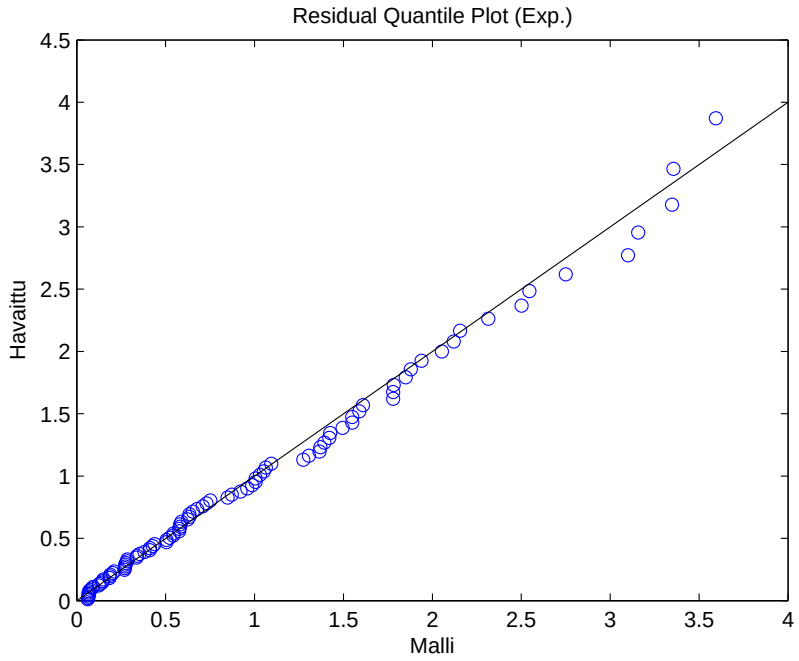
$$D = 2\left(l_2(\mathcal{M}_2) - l_0(\mathcal{M}_0)\right) = 2(-435.6 - (-440.2)) = 9.2.$$

Esimerkiksi merkitsevyystasolla $\alpha = 0.01$ kriittinen arvo on $c_{0.01} = 6.63 < D$, eli malli $\mathcal{M}_2$ on yksinkertaisempaa mallia tilastollisesti merkitsevästi parempi 99 %:n luottamustasolla (itse asiassa malli $\mathcal{M}_0$ voidaan hylätä aina 99.76 %:n luottamustasoon asti).

Tarkastellun, ajasta riippuvan skaalaparametrin sisältävän mallin log-uskottavuus on kuitenkin selvästi pienempi kuin trendin lokaatioparametrissä sisältävän mallin $\mathcal{M}_1$ edellisessä kohdassa: $l_2 = -435.6 < -406.6 = l_1$. Koska malleissa



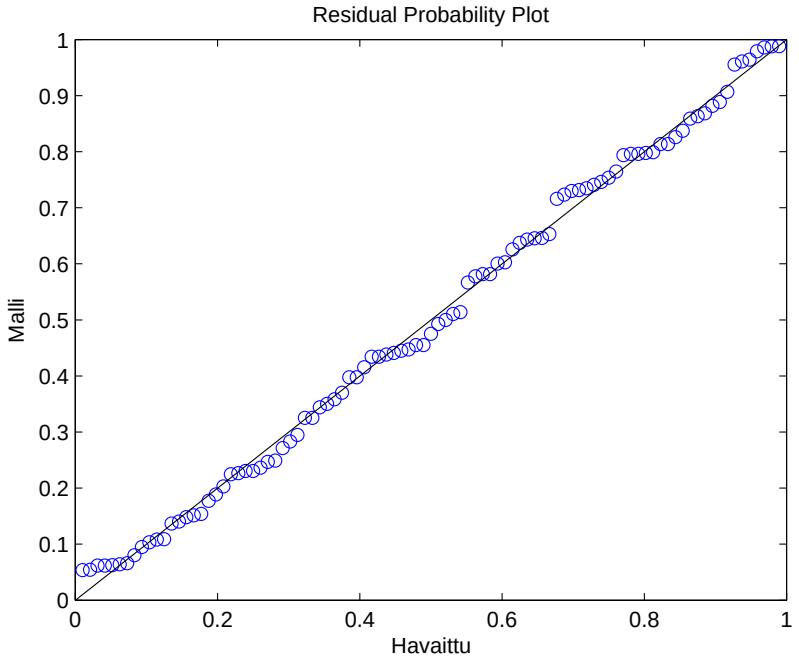
Kuva 2.48: Todennäköisyyskuvaaja mallille $\mathcal{M}_1$.



Kuva 2.49: Kvantiilikuvaaja mallille $\mathcal{M}_1$.

on yhtä monta parametria ja ne on estimoitu samasta datasta, voidaan loguskottavuuksia suoraan verrata ja päätyä johtopäätökseen, että malli $\mathcal{M}_1$ selittää havaittua dataa selvästi mallia $\mathcal{M}_2$ paremmin, vaikka molemmat ovat merkittäviä parannuksia perusmalliin $\mathcal{M}_0$.

Kuvissa 2.50 ja 2.51 on vielä todennäköisyys- ja kvantiilikuvaajat mallille. Sopivuus dataan vaikuttaa yleisesti ottaen hyvältä, mutta kvantiilikuvaajan perusteella ilmenee että jakauman hännässä yhteensopivuus ei ole niin hyvä kuin aiemmin.[15]



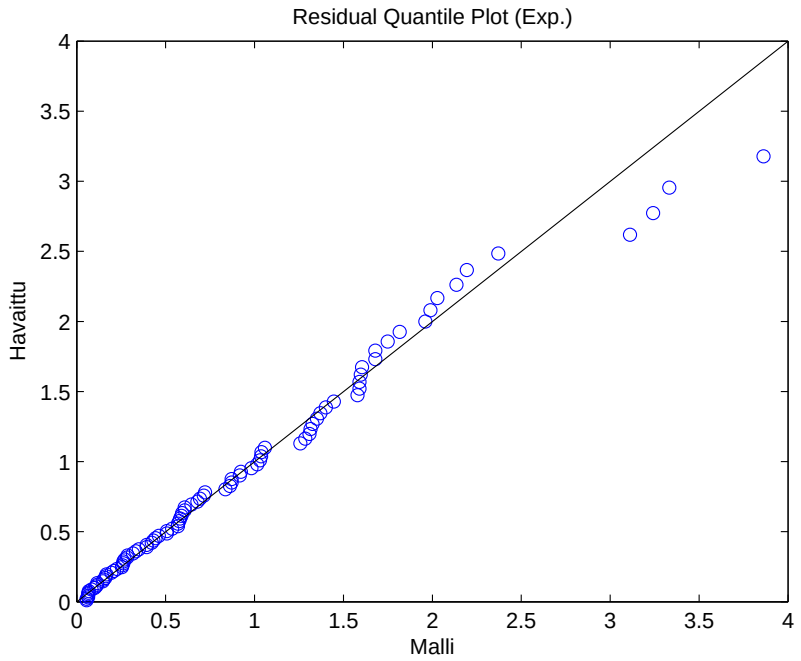
Kuva 2.50: Todennäköisyyskuvaaja mallille $\mathcal{M}_2$.

2.6.3.2 Selittävät muuttujat

Tarkastellaan seuraavaksi ulkoisten muuttujien (covariate) sisällyttämistä pisteprosessimalliin. Esimerkiksi monien luonnonilmiöiden kohdalla tarve tällaiselle on ilmeinen. Coles [1] mainitsee esimerkkinä saasteiden keskittymisen mallintamisen, jossa ilman saastepitoisuus riipuu yleensä vallitsevasta tuulen nopeudesta (kovalla tuulella on hajottava vaikutus saasteisiin). Samoin etenkin eteläisillä merillä vedenkorkeusmaksimit saattavat olla poikkeuksellisen korkeita niiden jaksojen aikoina, joina El Niño –ilmiö on aktiivinen.

Helsingin vedenkorkeuden kohdalla muistetaan, että tärkempiä Itämeren pinnankorkeuteen vaikuttavia tekijöitä ovat tuulet ja ilmanpaine. Pyritään mallintamaan meteorologisten olosuhteiden vaikutus vedenkorkeuteen käyttämällä

[15] Vrt. kohdan 2.6.3.3 esimerkkiin.



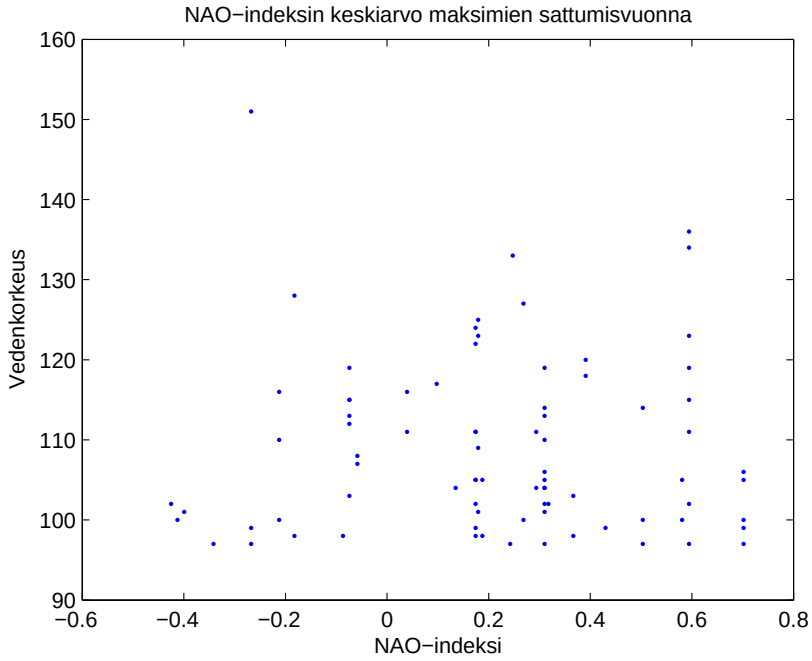
Kuva 2.51: Kvantiilikuvaaja mallille $\mathcal{M}_2$.

proxy-muuttujana NAO-indeksiä (North Atlantic Oscillation), joka ilmoittaa olennaisesti ilmanpaineen eron Islannin ja Azoreiden välillä, ja kuvaa länsituulien voimakkuutta. NAO-indeksin osalta on käytettävissä indeksin kuukausittaiset keskiarvot vuodesta 1950 lähtien.[16]

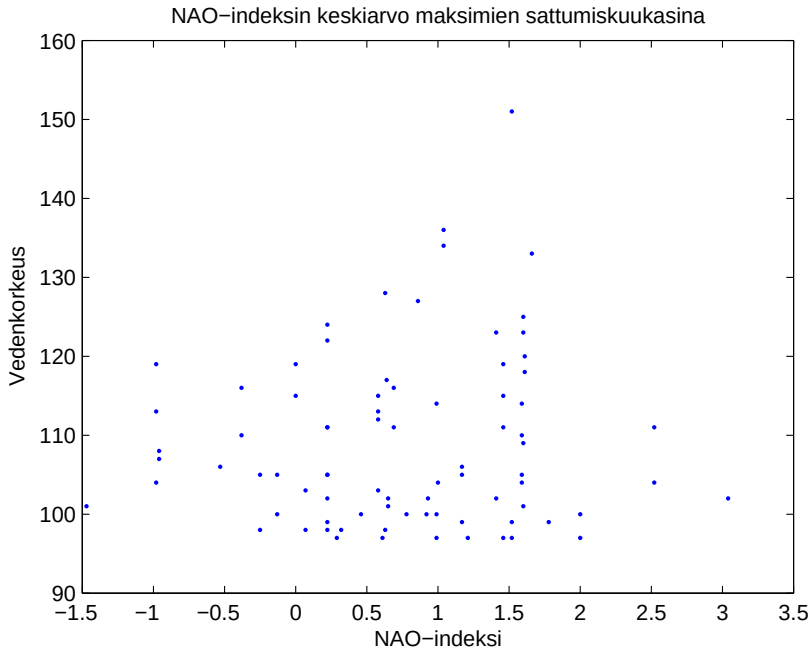
Kuvassa 2.52 on esitetty tason $u = 96$ cm ylittävät vedenkorkeushavainnot havainnon sattumisvuoden NAO-indeksin keskiarvon suhteen. Kuvan perusteella vaikuttaa siltä, että merenpinnan tasolla on taipumus olla korkealla silloin kun indeksin arvo on suuri. Vedenkorkeuden ja havaintovuoden NAO-indeksin vuosikeskiarvon välinen yksinkertainen lineaarinen korrelaatio on 0.407. Vedenkorkeuden ääri-ilmiöitä ajatellen vuosikeskiarvon käyttö saattaa kuitenkin olla liian karkeaa. Kuvassa 2.53 on esitetty edellistä kuvaa vastaavasti vedenkorkeus NAO-indeksin arvojen suhteen, mutta käyttäen kunkin vedenkorkeushavainnon havaintokuukauden keskiarvoa. Graafisesti vedenkorkeuden ja indeksin taipumus olla korkealla yhtä aikaa näyttää hieman vahvistuneen, vaikka lineaarinen korrelaatio (kuvan oikeassa laidassa sijaitsevista havainnoista johtuen) pysyykin käytännössä samana, ollen 0.412. Ääri-ilmiöiden mallintamista ajatellen käytetään kuitenkin NAO-indeksin sattumiskuukausien keskiarvoja, koska nämä luonnollisesti kuvaavat paremmin korkeiden merenpinnan tasojen havaitsemishetkillä vallinneita olosuhteita.

Lisätään edellisen perusteella NAO-indeksi Poisson-pisteprosessimalliin asetta-

[16]Lähde: (US) National Weather Service, Climate Prediction Center: `http://www-das.uwyo.edu/~geerts/cwx/notes/chap12/nao.html`.



Kuva 2.52: Tason $u = 96$ cm ylittävät vedenkorkeushavainnot piirrettynä sattumisvuoden NAO-indeksin keskiarvoa vasten.



Kuva 2.53: Tason $u = 96$ cm ylittävät vedenkorkeushavainnot piirrettynä sattumiskuukauden NAO-indeksin keskiarvoa vasten.

malla

$$\boldsymbol{\mu}(t) = \kappa_0 + \kappa_1 \mathrm{NAO}(t),$$

missä NAO(t) on NAO-indeksin arvo. Prosessin intensiteetiksi tulee siis

$$\lambda(t,x) = \frac{1}{\sigma}\left(1 + \xi\frac{x - (\kappa_0 + \kappa_1\mathrm{NAO}(t))}{\sigma}\right)^{-1/\xi - 1}$$

pisteeessä (t, x).

Käytettävä ylitedata koostuu nyt havainnoista 1.1.1950 alkaen, vastaten NAO-indeksin arvojen alkuhetkeä. Maksimoimalla log-uskottavuus saadaan parametriestimaateiksi

$$\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\kappa}_0, \hat{\kappa}_1, \hat{\sigma}) = (-0.224, 96.5, 12.6, 13.3),$$

ja log-uskottavuudeksi maksimissa -327.7.

Merkitään tarkasteltavaa mallia $\mathcal{M}_1^N$ ja verrataan tätä taas homogeeniseen POT-malliin. Jälkimmäinen täytyy nyt vertailtavuuden vuoksi estimoida uudestaan mallia $\mathcal{M}_1^N$ vastaavalta aikaperiodilta: merkitään näin saatua mallia $\tilde{\mathcal{M}}_0$. Perusmallin $\tilde{\mathcal{M}}_0$ log-uskottavuudeksi saadaan -405.3, ja testisuureen arvoksi tulee

$$D = 2\left(l_1(\mathcal{M}_1^N) - l_0(\tilde{\mathcal{M}}_0)\right) = 155.2,$$

mikä on erittäin suuri vastaavan χ^2-jakauman kvantiileihin verrattuna, ja implikoi NAO-indeksin sisältävän mallin olevan perusmallia parempi ilmiön eli merenpinnan korkeuden ääriarvojen kuvaamisessa. Vertailun vuoksi aikavälin 1950-2011 ylitehavainnoista estimoitujen mallien $\mathcal{M}_1$ (μ-trendi) ja $\mathcal{M}_2$ (σ-trendi) maksimoidut log-uskottavuudet ovat -344.8 ja -357.9.

Todennäköisyys- ja kvantiilikuvaajat mallille (kuvat 2.54 ja 2.55) vahvistavat mallin hyvän sopivuuden havaintoihin.

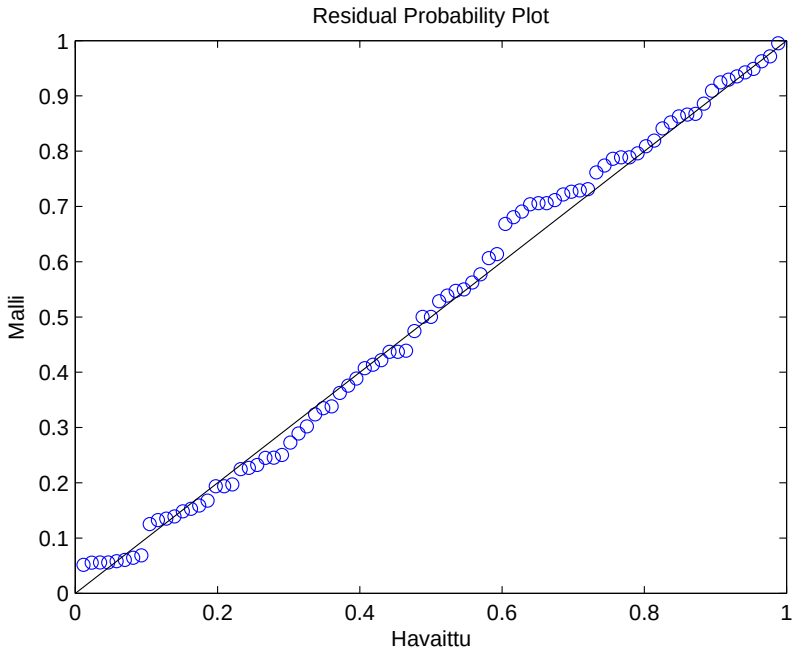
Vaikka NAO-indeksin lisääminen malliin parantaa mallisovitetta suuresti, pelkästään edellisen analyysin perusteella ei voida vielä sanoa mitään kovin varmaa vedenkorkeuden ja NAO:n riippuvuussuhteesta. Edellisessä osiossa (ja jo aiemmin) nähtiin, että vedenkorkeusmaksimeissa on havaittavissa vahvaa evidenssiä yleisestä nousevasta trendistä. On siis mahdollista, että myös NAO-indeksin arvot ovat ajassa muuttuvia, ja suurten vedenkorkeushavaintojen sekä NAO-indeksin arvojen välillä näyttäytyvä riippuvuus on vain seurausta niiden molempien muutoksesta ajassa. Tämän tarkastelemiseksi laajennetaan mallia ottamalla mukaan myös aikatrendi, ja asetetaan

$$\boldsymbol{\mu}(t) = \kappa_0 + \kappa_1 t + \kappa_2 \mathrm{NAO}(t).$$

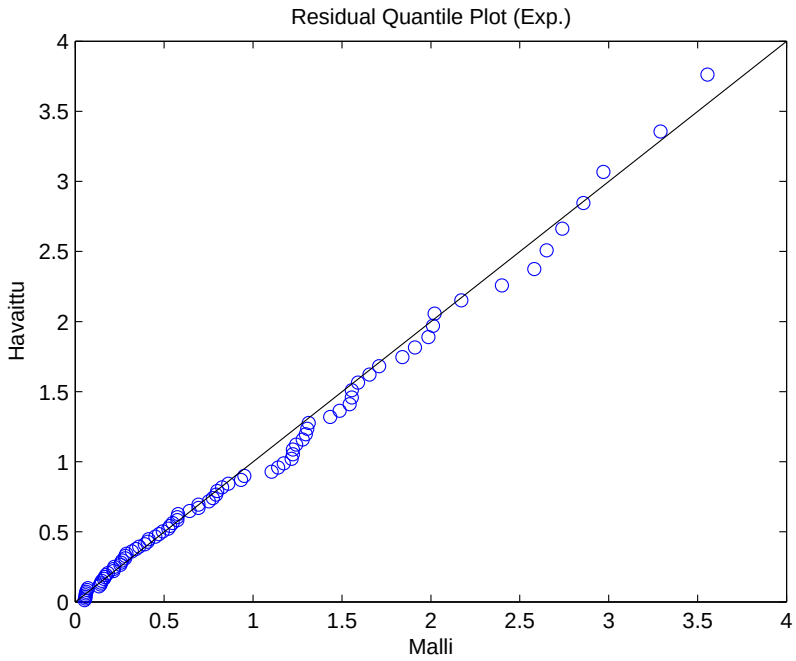
Merkitään näin saatua mallia $\mathcal{M}_2^N$. Sovittamalla malli dataan saadaan parametrien SU-estimaateiksi

$$\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\kappa}_0, \hat{\kappa}_1, \hat{\kappa}_2, \hat{\sigma}) = (-0.202, 83.8, 0.37, 11.4, 12.7),$$

ja log-uskottavuudeksi -319.8. Parametrin κ_1 estimaatin arvo tulkitaan siten, että mallin mukaan vedenkorkeus*maksimeissa* on kasvava trendi, kasvun ollessa keskimäärin n. 0.37 cm vuodessa. Estimaatin arvo vastaa melko tarkkaan pelkän



Kuva 2.54: Todennäköisyyskuvaaja mallille $\mathcal{M}_1^N$.



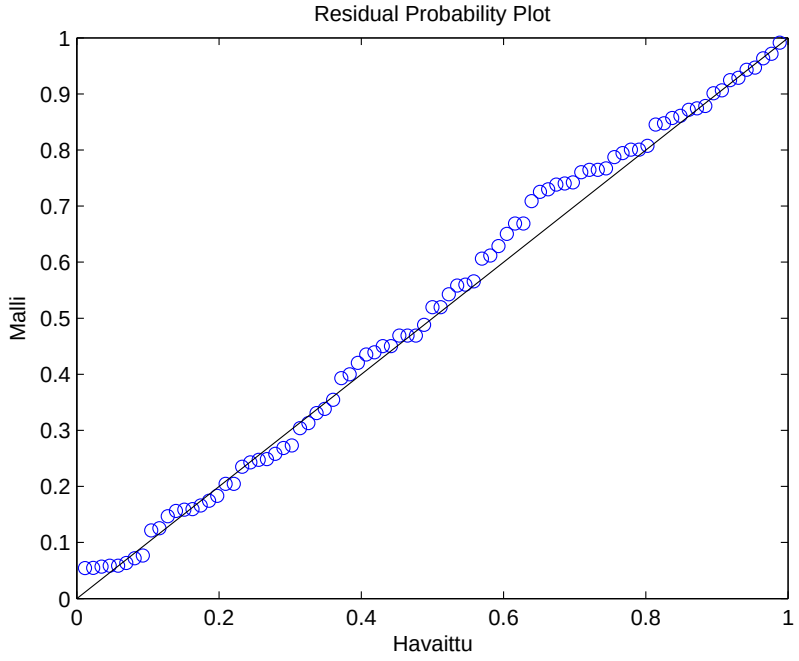
Kuva 2.55: Kvantiilikuvaaja mallille $\mathcal{M}_1^N$.

trendin sisältävän mallin $\mathcal{M}_1$ vastaavaa, 0.39 cm. NAO-indeksin lisääminen malliin ei siis näytä käytännössä muuttavan tätä arviota. Sen sijaan aikatrendin lisääminen pelkän NAO:n sisältävään malliin $\mathcal{M}_1^N$ laskee parametrin κ_2 estimaatin arvoa 12.6 cm:stä yllä olevaan 11.4 cm:iin. Parametri κ_2 voidaan tulkita siten, että NAO-indeksin arvon kasvaessa yhden yksikön, kasvavat vedenkorkeuden äärihavaintojen suuruudet keskimäärin 11.4 cm.

Verrataan mallia nyt edelliseen malliin $\mathcal{M}_1^N$, joka siis on mallin $\mathcal{M}_2^N$ erikoistapaus, kun $\kappa_1 = 0$ yllä. Testisuureen arvoksi saadaan

$$D = 2\left(l_2(\mathcal{M}_2^N) - l_1(\mathcal{M}_1^N)\right) = 15.8,$$

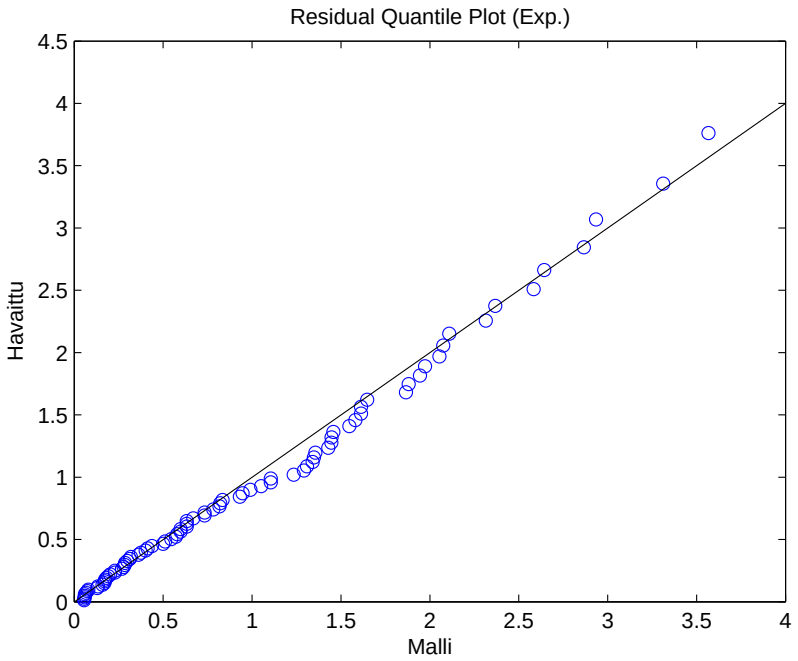
mikä on jälleen suuri, osoittaen että datan sisältämä todistusaineisto mallin $\mathcal{M}_2^N$, puolesta mallia $\mathcal{M}_1^N$ vastaan on hyvin vahvaa. Kuvissa 2.56 ja 2.57 on esitetty vielä todennäköisyys- ja kvantiilikuvaajat mallille. Nämä eivät aseta mallin sopivuutta havaintoihin kyseenalaiseksi.



Kuva 2.56: Todennäköisyyskuvaaja mallille $\mathcal{M}_2^N$.

2.6.3.3 Alin rakentamiskorkeus rannikolla

Sovelletaan edellä saatuja malleja Helsingin rannikolla vaadittavan rakentamiskorkeuden arvioimiseen. Suomen Ympäristökeskuksen, ympäristöministeriön sekä maa- ja metsätalousministeriön suosituksessa alimmalla hyväksyttävällä rakentamiskorkeudella tarkoitetaan ylintä korkeutta, jolle vesi voi nousta ilman, että se vahingoittaa rakennuksen rakenteita [32, luku 6]. Alin tulvien kannalta hyväksyttävä rakentamiskorkeus määritetään meren rannikolla olennaisesti



Kuva 2.57: Kvantiilikuvaaja mallille $\mathcal{M}_2^N$.

perustuen korkeuteen, jonka tulva ylittää keskimäärin kerran 200 vuoden aikana.[17] Sisävesistöissä vaatimus on asetettu tulvakorkeudeksi, joka ylittyy kerran 50 vuodessa.[18]

Helsingissä alimmat rakennuskorkeudet määritetään oppaan [32] mukaisesti. Oppaassa [32] mainittu Merentutkimuslaitoksen arvio keskimäärin kerran seuraavassa 200 vuodessa saavutettavalle vedenpinnan korkeudelle Helsingin rannikolla on 236 cm. Helsingin tulvastrategian mukaan "on huomattava, että tässä arviossa otetaan huomioon esimerkiksi IPCC:n kolmannesta arviointiraportista poiketen Grönlannin mahdollisen mannerjään sulamisen kiihtymisen vaikutus valtamerien pinnankorkeuteen ja sen vaikutus edelleen Itämereen". [32, s. 6]

Verrataan edellisten osioiden mallien antamia tuloksia edellä mainittuun, eli estimoidaan vedenpinnan korkeus, joka ylitetään keskimäärin kerran seuraavan

[17] Tähän lisätään harkinnanvarainen lisäkorkeus vähintään n. 0.3–1.0 metriä, ja avointen ulapoiden rannikoilla harkinnanvarainen aaltoiluvara.

[18] Lisäksi suomalaisilla vahinkovakuutusyhtiöillä on tavallisesti kotivakuutukseen sisältyvässä luonnonilmiöturvassa korvauspiirin rajoitus, jonka mukaan vakuutuksesta korvataan vain *poikkeuksellinen* vesistö- tai merivesitulva. Poikkeuksellisena vedenpinnan tai merenpinnan nousuna pidetään ehdoissa yleensä vedenkorkeutta, jonka esiintymistodennäköisyys on kerran 50 vuodessa tai harvemmin.

200 vuoden aikana. Osion 2.5.3 tapaan taso x_m saadaan nyt yhtälöstä

$$\left(1-\frac{1}{m}\right)^n = \mathbb{P}(\max X_1,\ldots,X_n \le x_m) = \prod_{t=1}^{n} H_{\hat{\boldsymbol{\theta}}(t)}(x_m)$$
$$= \prod_{t=1}^{n} \exp\left\{-\left(1+\hat{\boldsymbol{\xi}}(t)\frac{x_m-\hat{\boldsymbol{\mu}}(t)}{\hat{\boldsymbol{\sigma}}(t)}\right)^{-1/\hat{\boldsymbol{\xi}}(t)}\right\},$$

missä $m = 200$, $n = 200$, ja $t = 0$ vastaa vuotta 2012. Taso x_m täytyy ratkaista numeerisesti: tätä varten on mukavampaa saattaa yhtälö muotoon

$$\ln\left(1-\frac{1}{m}\right) = -\sum_{t=1}^{n}\frac{1}{n}\left(1+\hat{\boldsymbol{\xi}}(t)\frac{x_m-\hat{\boldsymbol{\mu}}(t)}{\hat{\boldsymbol{\sigma}}(t)}\right)^{-1/\hat{\boldsymbol{\xi}}(t)}.$$

Vaihtoehtoisesti voidaan tarkastella ongelmanasettelun sanamuodon mukaisesti suoraan tason x_m ylityksien eli tapahtumien $\{X_i > x_m\}$ lukumäärän (merk. N_m) odotusarvoa:

$$\mathbb{E}(N_m) = \mathbb{E}\left(\sum_{i=1}^{n}\mathbb{1}_{\{X_i>x_m\}}\right) = \sum_{i=1}^{n}\mathbb{P}(X_i > x_m)$$
$$= \sum_{i=1}^{n}\big(1-\mathbb{P}(X_i \le x_m)\big) = \sum_{i=1}^{n}\left(1-H_{\hat{\boldsymbol{\theta}}(t)}(x_m)\right).$$

Odotusarvoisesti kerran seuraavassa m vuodessa ylitettävä taso saadaan, kun asetetaan yhtälö ykköseksi ja ratkaistaan x_m. Molemmat laskutavat antavat luonnollisesti identtisen vastauksen.

Trendin lokaatioparametrissä μ sisältävä malli $\mathcal{M}_1$ antaa korkeudeksi $x_m = 221$ cm. Malli $\mathcal{M}_2$, joka sisältää trendin skaalaparametrissä σ, taas antaa vedenkorkeudeksi 386 cm. Osoittautuu siis, että vaikka malli $\mathcal{M}_2$ sopi havaittuihin vedenkorkeuksiin kohtuullisen hyvin (erityisesti sen saavuttamalla loguskottavuudella mitattuna), mallissa käytetty lineaarinen trendi jakauman skaalaparametrin logaritmimuunnoksessa tuottaa epäuskottavan suuren vedenkorkeuden arvon pitkälle eteenpäin projektoidessa. Vertaa myös kvantiilikuvaajaan 2.51, joka viittaa huonohkoon sopivuuteen juuri jakauman oikeassa hännässä.

Puhtaasti tilastollisen mallin $\mathcal{M}_1$ antama vedenkorkeuden estimaatti 221 cm on jonkin verran pienempi mutta karkeasti samaa suuruusluokkaa Merentutkimuslaitoksen arvion 236 cm kanssa. Tarkastellusta esimerkistä käy selväksi, että vedenkorkeuden ääriarvoissa havaittu muutos on ensiarvoisen tärkeää huomioida pidemmän aikavälin ennusteita tarkastellessa. Esimerkki myös havainnollistaa niitä vaaroja, jotka liittyvät kohtuullisen hyvinkin havaintoihin sopivan mallin suoraviivaiseen ekstrapolointiin, jos sopivuus sovelluksen kannalta kriittisellä alueella – tässä jakauman äärimmäisessä oikeassa hännässä – on vähemmän tyydyttävä.

Mikäli NAO-indeksin kehitystä arvioidaan tai tulevia arvoja simuloidaan, voidaan indeksin sisältävää mallia $\mathcal{M}_2^N$ vastaavalla tavalla käyttää kerran seuraavassa m. vuodessa esiintyvän tulvakorkeuden estimointiin. Simuloimalla NAO-indeksin kehitystä ja laskemalla haluttu suure jokaiselta simulaatiopolulta, saa-

daan suureelle (esim. kerran 200 vuodessa sattuvalle vedenkorkeudelle) muodostettua todennäköisyysjakauma, kun simulaatio toistetaan riittävän monta kertaa. Estimaattiin liittyvän epävarmuuden arviointi simuloidun todennäköisyysjakauman perusteella on suoraviivaista.

Luku 3

Katastrofikuolemien määrän arvioinnista

Vakuutusyhtiölain (2008/521) 1 L 16 §:n mukaan "[v]akuutusyhtiön toimintapääoma, jälleenvakuutus ja muut yhtiön vakavaraisuuteen vaikuttavat seikat on järjestettävä vakuutetut edut turvaavalla tavalla ottaen huomioon tuottojen ja kulujen todennäköinen vaihtelu sekä arvioitavissa olevat muut epävarmuustekijät" (ns. turvaavuusperiaate). Tämä vaatii siis muun ohessa jälleenvakuutustarpeen eksplisiittistä arviointia ja tarvittaessa jälleenvakuutuksen järjestämistä asianmukaisella tavalla. Erityisen tärkeää jälleenvakuutustarpeen oikea arvioiminen on sellaisten katastrofaalisten vahinkojen osalta, joiden todennäköisyys kyllä saattaa olla pieni, mutta jotka toteutuessaan vakavasti heikentävät yhtiön vastuunkantokykyä, tai jopa aiheuttavat yhtiön vararikon. Käänteisesti, jälleenvakuuttajan keskeisenä ongelmana on ensivakuuttajalle myönnetyn katastrofisuojan oikea hinnoittelu.

Solvenssi II –puitedirektiivin[1] mukaan vakuutusyhtiön perusvakavaraisuuspääomavaatimuksen laskennassa (standardikaavalla tai sisäisellä mallilla) on otettava yhtenä osana huomioon pääomavaatimus joka syntyy henki-, vahinko- tai sairausvakuutukseen liittyvästä katastrofiriskistä; jälkimmäinen on direktiivin mukaan "tappioriski tai vakuutusvelkojen arvossa tapahtuvan epäedullisen muutoksen riski, joka johtuu merkittävästä epävarmuudesta hinnoittelua ja vastuuvelkaa koskevissa oletuksissa, jotka liittyvät äärimmäisiin tai poikkeuksellisiin tapahtumiin".

Tarkastellaan seuraavassa henkivakuutustyyppiseen toimintaan liittyvää katastrofiriskiä. Tavoitteena on tutkia, mitä ääriarvoteoriaa soveltamalla voidaan sanoa suomalaisia koskevien katastrofikuolemien todennäköisyysjakaumasta. "Katastrofin" voisi ehkä korvata tässä yhteydessä "onnettomuudella", sillä Suomessa ei ole (onneksi!) sattunut juurikaan sen mittakaavan onnettomuuksia, mitä katastrofilla yleisessä kielenkäytössä yleensä ymmärretään. Tämä luonnehtiikin toista analyysin kannalta merkittävistä piirteistä: dataa suomalaisia kohdan-

[1] Euroopan parlamentin ja neuvoston direktiivi 2009/138/EY vakuutus- ja jälleenvakuutustoiminnan aloittamisesta ja harjoittamisesta (Solvenssi II).

neista suuronnettomuuksista on verraten niukasti, varsinaisista katastrofeista ei käytännössä ollenkaan.

Toisaalta tiedetään, että jättimäiset katastrofitkin ovat mahdollisia, myös Suomessa: esimerkiksi Tunguskan yllä vuonna 1908 räjähtänyt asteroidi olisi muutamia tunteja myöhemmin (ja vain hiukan eri leveysasteella) ilmakehään syöksyessään osunut maapallon pyörimisliikkeen vuoksi Helsinkiin, tuhoten sen lähialueineen täydellisesti. Periaatteessa katastrofien suuruudella ei myöskään ole olemassa ylärajaa, mitattiin suuruutta sitten kuolemien lukumäärällä, tai jollain muulla tavalla.[2] On siis mahdollista rakentaa skenaarioita, joissa katastrofi on miten suuri tahansa (ja joskus jopa arvioida karkeasti näiden todennäköisyyksiä suuruusluokkatasolla), mutta käytännön vakuutustoiminnan kannalta tämä ei ole erityisen hyödyllistä.

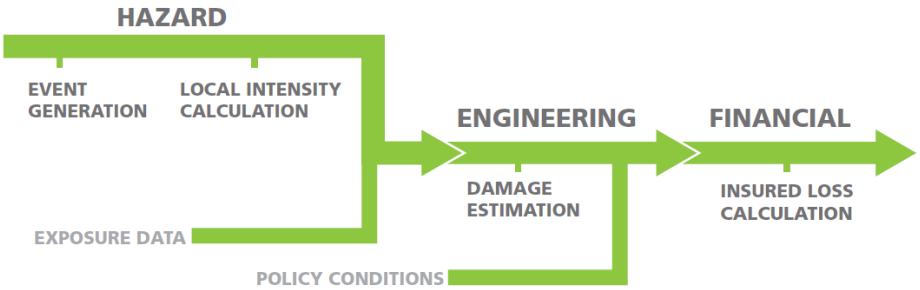
Toisaalta katastrofin vaikutus yksittäiseen vakuutusyhtiöön – kun vaikutukset finanssimarkkinoihin ja sitä kautta yhtiön varoihin ja vakavaraisuuteen rajataan tässä yhteydessä tarkastelun ulkopuolelle – rajoittuu korvausvelvollisuuteen, joka yhtiölle katastrofin seurauksena syntyy olemassa olevien vakuutussopimusten ja mahdollisen otetun jälleenvakuutusliikkeen perusteella. Sopimusten perusteella syntyvä maksimitappio on siis yleensä tiedossa. Tämän maksimitappion todennäköisyyden arviointi on jo käytännössäkin kiinnostavaa. Vahinko- ja jälleenvakuutuksessa käytetään yleisesti termiä Estimated Maximum Loss (EML) tai Probable Maximum Loss (PML) kuvaamaan tietystä vakuutetusta kohteesta tai sopimuksesta aiheutuvaa suurinta "odotettavissa olevaa" tappiota (esim. [16]) – tämä määritellään käytännössä yleensä jakauman hännän prosenttipisteenä[3].

Katastrofien aiheuttamien kokonaistappioiden jakauma on siis yleisessä tapauksessa erilainen kuin kyseisistä katastrofeista tietylle yhtiölle aiheutuvien tappioiden jakauma. Käytännössä vakuutusyhtiöön kohdistuvaa katastrofiriskiä kvantifioidaan arvioimalla ensin katastrofin aiheuttavien tapahtumien todennäköisyyttä ja suuruutta (ja tarvittaessa maantieteellistä sijaintia), ja sitten tällaisten tapahtumien vaikutusta vakuutusyhtiöön yhtiön myöntämien sopimusten kautta. Tällainen arviointi on usein luonnollista toteuttaa simulaation keinoin. Kuvassa 3.1 on esimerkki katastrofimallinnukseen erikoistuneen AIR Worldwiden esitteestä[4]: Ensimmäisessä vaiheessa simuloidaan suuri joukko mahdollisia tapahtumia ja lasketaan tapahtumien voimakkuus niiden vaikutuspiiriin kuuluvan alueen eri pisteissä. Seuraavaksi tähän yhdistetään data ilmiölle altistuneista kohteista (esimerkiksi simuloidun maanjäristyksen vaikutusalueeseen kuuluvista rakennuksista), ja määritetään kohteille aiheutuneet vahingot, jotka edelleen muutetaan rahasummiksi. Lopulta lasketuista kokonaisvahingoista saadaan vakuutusten piiriin kuuluvat korvattavat vahingot soveltamalla vahingolle altistuneiden vakuutussopimusten sopimusehtoja. Kun simulaatioita suoritetaan riittävän suuri määrä, saadaan käsitys vahinkojen todennäköisyysjakaumasta eli vahinkojakaumasta (käytettäviin oletuksiin perustuen).

[2] Esimerkiksi riittävän lähellä (kosmisessa mittakaavassa) esiintyvä supernova eli massiivisen tähden räjähdys voisi tuhota kaiken elämän maapallolta. Ks. esim. Wikipedia, `http://en.wikipedia.org/wiki/Supernova`.

[3] Samasta kvantiilipohjaisesta riskimitasta käytetään finanssialalla nimitystä Value-at-Risk, ks. luku 4.

[4] Esitys on löydettävissä osoitteesta `http://www.air-worldwide.com/Models/Overview/`.



Kuva 3.1: Katastrofitappioiden mallinnuksen viitekehys (ks. alaviite 4).

Kuvattua lähestymistapaa käytetään pääsääntöisesti vakuutetulle omaisuudelle aiheutuvien vahinkojen arvioimiseen. Samoja yleisperiaatteita voidaan kuitenkin periaatteessa soveltaa myös henkilövahinkoihin.[5] Lähestymistapa toimii erityisen hyvin, kun tarkastellaan tietyn, hyvin määritellyn ilmiön aiheuttamia vahinkoja. Tällaisia ovat luonnonilmiöt – kuten maanjäristykset, hurrikaanit, maastopalot, tulvat jne. – joiden taustalla on selkeä fysikaalinen prosessi, ja joiden esiintymistä ja vaikutuksia voidaan kohtalaisella tai hyvällä menestyksellä mallintaa ilmiöiden luonnetta koskevaan tieteelliseen tutkimustietoon perustuen.

Riskihenkivakuutusten osalta tilanne on monimutkaisempi. Toinen tämän analyysin kannalta merkittävä piirre datan niukkuuden lisäksi onkin tarkasteltavan aineiston taustalla olevan yksikäsitteisen fysikaalisen prosessin puuttuminen – tai oikeammin, se että taustalla oleva prosessi on monien yksittäisten (toisistaan enemmän tai vähemmän riippuvien tai riippumattomien) prosessien yhdistelmä, koska henkivakuutusyhtiöiden myöntämien henki- eli kuolemanvaraturvien korvauspiiriin kuuluvat lähtökohtaisesti kaikki kuolemansyyt.[6] Samoin henki- ja vahinkovakuutusyhtiöiden myöntämistä tapaturmaisen kuoleman varalta voimassa olevista vakuutuksista korvataan kaikki sellaisista tapaturmista johtuvat kuolemat, joita ei ole erikseen vakuutusehdoissa rajattu vakuutusturvan ulkopuolelle.

Mahdollisten kuolinsyiden lukuisuus tekee edellä kuvatunlaisen kausaalisen lähestymistavan soveltamisen vaikeaksi. Vaihtoehtona on arvioida katastrofikuolemien määrää puhtaasti tilastollisesti kuolemien syitä mallintamatta. Tämä on tavoitteena tässä luvussa. Mallinnuksen tuloksena saadaan katastrofikuolemien lukumäärän todennäköisyysjakauma, jonka perusteella voidaan arvioida erisuuruisten tapahtumien todennäköisyyttä, ja jota käyttäen voidaan simuloida kata-

[5]Tosin henkilövahinkoja tarkastellessa (simuloitujen) katastrofien vaikutusta harvoin pystytään määrittämään niin tarkasti kuin vakuutettujen rakennusten ja muun kiinteän omaisuuden kohdalla, joiden maantieteellinen sijainti on täsmällisesti tunnettu ja pysyvä, ja ominaisuudet – kuten rakenne ja materiaalit – tarkkaan tiedossa.

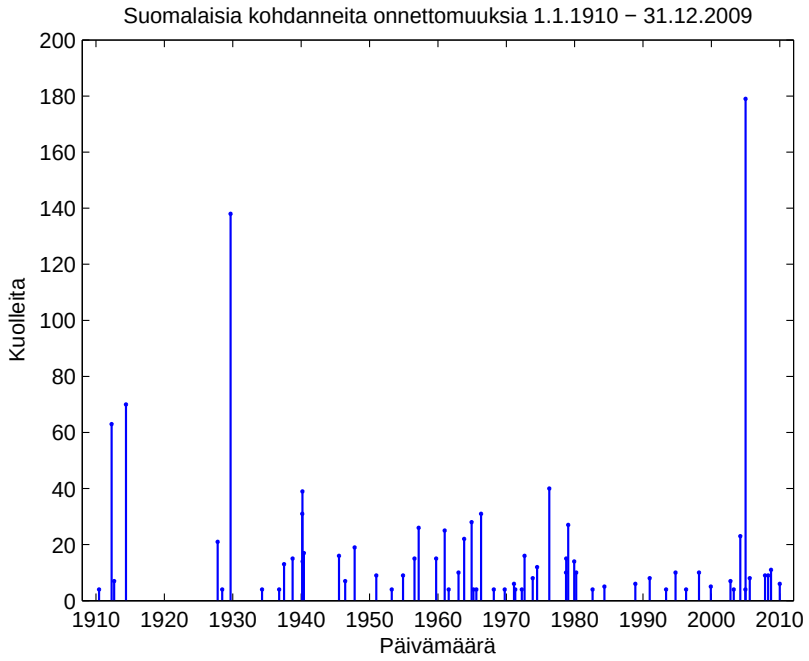
[6]Kuolemantapaussummaa ei makseta, jos vakuutettu on tehnyt itsemurhan ennen kuin vuosi on kulunut vakuutusyhtiön vastuun alkamisesta. Korvausta ei myöskään säännönmukaisesti makseta, jos vakuutetun kuolema on aiheutunut osallistumisesta ulkomailla tapahtuvaan sotaan tai aseelliseen selkkaukseen (pl. rauhanturvaamistehtävät tietyin ehdoin), tai jos vakuutettu on kuollut ihmisiä joukoittain tuhonneen ydinreaktioon perustuvan aseen tai laitteen äkillisestä vaikutuksesta. Vakuutusyhtiön vastuusta Suomen joutuessa sotaan tai aseelliseen selkkaukseen on säädetty laissa erikseen (Laki henkivakuutuksesta sodan aikana, 364/1939).

strofikuolemien lukumääriä. Tarkastelu koskee siis kuvan 3.1 ensimmäistä osiota (Hazard). Tapahtumien vaikutusta määrättyyn vakuutusyhtiöön voidaan arvioida tämän perusteella ottamalla huomioon kannan rakenne ja sopimusten luonne (exposure data ja policy conditions ed. kuvassa).

3.1 Aineisto

Tarkastelun pohjana oleva aineisto koostuu sellaisista suomalaisia kohdanneista onnettomuuksista vuosilta 1910–2009, joissa on kuollut yli kolme henkilöä.[7] Tämä raja päädyttiin asettamaan onnettomuuksissa kuolleiden lukumäärien tutkimisen perusteella ja mahdollisimman täydellisen havaintoaineiston saamiseksi, sillä suuremmat tapaukset on paremmin dokumentoitu. Tällä rajauksella ei ole käytännössä vaikutusta analyysin lopputulokseen, sillä tarkastelussa ollaan kiinnostuneita selvästi tätä tasoa suuremmista tappioista.

Kustakin onnettomuudesta analysoitavaan dataan otettiin onnettomuuden tapahtumapäivä ja siinä kuolleiden henkilöiden lukumäärä. Näiden muodostama havaintosarja on esitetty kuvassa 3.2. Aineistosta on jätetty pois ilmiselvästi sotatoimiin liittyvät onnettomuuskuolemat, sillä näitä voidaan pitää käsiteltävän sovelluksen kannalta tuloksia vääristävinä.



Kuva 3.2: Suomalaisia kohdanneita suuronnettomuuksia päivämäärän mukaan.

[7] Data perustuu Tapani Tuomiselta (Keskinäinen Vakuutusyhtiö Kaleva) saatuun, julkisista lähteistä kerättyyn aineistoon; ks. erityisesti suomenkielinen Wikipedia (fi.wikipedia.org), luokka: "Onnettomuudet Suomessa".

Taulukossa 3.1 on esitetty joitakin tavanomaisia tilastollisia tunnuslukuja havaintoaineistolle. Yli kolme henkeä vaatineita onnettomuuksia on tarkastelujaksolla sattunut 65 kpl. Mukana on lukuisia alarajalle osuvia havaintoja eli 4 henkeä vaatineita onnettomuuksia, kuten moodin arvo taulukossa osoittaa; suurin tarkastelujaksolla sattunut onnettomuus vaati 179 henkeä (Tapaninpäivän tsunami vuonna 2004). Kaksi suurinta havaintoa erottuu koollaan muista, ja kaksi seuraavaa edelleen jäljelle jäävistä.

Taulukko 3.1: Tilastollisia tunnuslukuja suomalaisten onnettomuuskuolemadatalle.

n	min	max	mediaani	moodi	keskiarvo	keskihajonta	IQR
65	4	179	10	4	18.2	28.6	13.5

Kerätty aineisto kattaa pitkän, 100 vuoden aikajakson. Kuten aina pitkiä havaintoaikasarjoja tarkastellessa, herää kysymys siitä, ovatko eri aikojen havainnot yhteismitallisia – tai, ovatko taustalla olevan ilmiön tilastolliset ominaisuudet pysyneet muuttumattomina. Esimerkiksi rahamääräisiä suureita tarkasteltaessa pitkällä aikavälillä on välttämätöntä huomioida rahan arvon muutos eli inflaatio. Onnettomuuskuolemien kohdalla asia ei ole yksiselitteinen. Datan muokkaamistarpeen arvioimiseksi tarkastellaan seuraavaksi hieman (suur)onnettomuuksien aiheuttamia kuolemia ilmiönä.

Kun tarkastellaan yhteiskuntia laajasti, onnettomuuskuolemien sattumistiheyteen ja suuruuteen vaikuttavia pääasiallisia seikkoja voidaan ajatella olevan populaation koko, taloudellisen aktiviteetin taso ja ilmenemismuodot, sekä maantieteellinen sijainti. Tuntuu selvältä, että populaation koko vaikuttaa onnettomuuskuolemien sattumiseen sekä suoraan että välillisesti taloudellisen aktiviteetin kautta: mitä enemmän ihmisiä ja toimintaa, sitä enemmän mahdollisuuksia onnettomuuksien sattumiseen, *ceteris paribus*. Toisaalta tietynlaisen onnettomuuden sattuessa sen aiheuttama tuho voi olla suurempi väkiluvultaan suuremmassa populaatiossa, koska onnettomuudelle altistuneita on potentiaalisesti enemmän. Edelleen taloudellinen aktiviteetti ja sen ilmenemismuodot voivat vaikuttaa siten, että tietynlaisessa populaatiossa suuronnettomuuksien riski voi olla suurempi kuin toisessa, vaikka jälkimmäinen populaatio olisi kooltaan suurempi. Esimerkkinä voidaan ajatella vaikkapa maata, jossa harjoitetaan paljon merenkulkuun liittyvää elinkeinotoimintaa – tai vaikkapa raskasta teollisuustoimintaa vanhentuneella teknologialla – verrattuna maahan, jossa harjoitetaan henkilöihin kohdistuvalta suuronnettomuusriskiltään pienempää maanviljelystä. Elinkeinorakenne siis vaikuttaa (suur)onnettomuusriskiin. Kolmanneksi, maantieteellisellä sijainnilla on olennainen vaikutus sen kannalta, millaisten luonnonkatastrofien sattuminen on mahdollista ja todennäköistä. Esimerkiksi hurrikaanit, maanjäristykset ja taifuunit ovat merkittävimpiä katastrofien aiheuttajia niillä alueilla, joilla niitä esiintyy.

Palaten käsillä olevaan aineistoon, Suomen populaatio on kasvanut tarkastelujakson aikana enemmän tai vähemmän tasaisesti vuoden 1910 2.934 miljoonasta vuoden 2009 5.351 miljoonaan. Yksi mahdollisuus olisi skaalata tai inflatoida aineiston havaitut katastrofikuolemien lukumäärät vastaamaan esimerkiksi tarkastelujakson viimeisen eli vuoden 2009 lopun väkilukua. Aineistossa (kuva 3.2) ei kuitenkaan ole alustavalla tutkimisella havaittavissa trendiä tai muuta syste-

maattista muutosta, ei suuronnettomuuksien sattumisfrekvenssissä eikä niiden suuruudessa. Tämä siitä huolimatta, että Suomen väkiluku on kasvanut n. 80 prosenttia tarkastelujakson aikana, ja väkiluvun lisäksi sekä seurauksena myös taloudellinen aktiviteetti on kasvanut voimakkaasti.

Tämän esityksen tavoitteena ei ole analysoida suuronnettomuuksien sattumiseen vaikuttavia syitä sinänsä, sen enempää kuin aineiston käsittely tilastollista mallinnusta varten edellyttää. Mahdollinen selitys havaitulle trendin poissaololle suuronnettomuuksien sattumistiheyden osalta on kuitenkin intuitiivisesti teknologian kehittyminen ja sitä kautta turvallisuuden lisääntyminen. Tämä vähentää onnettomuuksien sattumismahdollisuutta, kun taas väestön kasvaminen ja aktiviteetin lisääntyminen lisäävät sitä. Suuronnettomuuksien suuruuksien taustoittamiseksi täytyy ajatella onnettomuuksien syntytapaa: useimmat havaintoaineiston suuremmat onnettomuudet koskevat liikennettä, oli kyseessä sitten maantie-, juna-, lento- tai meriliikenne. Tällöin onnettomuuksille altistuneiden lukumäärän ratkaisee luonnollisesti liikennevälineen kantokapasiteetti, joka ei juuri ole muuttunut tarkastelujakson aikana. Esimerkiksi juniin on 1900-luvun alkupuolella mahtunut karkeasti samaa suuruusluokkaa oleva henkilömäärä kuin 2000-luvun alkupuolella; väestönkasvun ei täten odoteta näkyvän yksittäisissä junaonnettomuuksissa kuolleiden lukumäärien kasvuna, vaan junaliikenteen lisääntymisen kautta onnettomuuksien lukumäärän kasvuna; teknologian kehitys ja raideliikenteen turvallisuuden parantuminen kuitenkin vaikuttaa samaan aikaan tasapainottavasti onnettomuuksien lukumäärää vähentäen. Samanlainen päättely pätee myös moneen muuhun onnettomuustyyppiin.

Edellä esitetyn intuitioon nojaavan päättelyn perusteella havaintoaineistoa ei skaalata tai muokata, vaan se otetaan sellaisenaan tilastollisen analyysin lähtökohdaksi.

3.2 Onnettomuuskuolemien mallintaminen

Koska havaintoaineisto sisältää onnettomuudet, joissa on kuollut yli 3 henkilöä, on se valmiiksi ylitedatan muodossa kynnyksellä $u = 3$. Täten on luonnollista aloittaa onnettomuuskuolemien suuruuksien tarkastelu soveltamalla ylitemenetelmää. Datasta myös puuttuu struktuuri, toisin kuin oli esimerkiksi vedenkorkeusdatan kohdalla, jossa data muodostui tasavälisistä havainnoista päivittäisten maksimien muodossa (ja näistä voitiin muodostaa esimerkiksi vuosimaksimien aikasarja). Tämä struktuurin puute tekee vuosi-/blokkimaksimimenetelmästä soveltumattoman tarkasteltavaan dataan.

Luodaan katsaus onnettomuuskuolemadataan olettaen ensin ainoastaan, että havaintojen jakaumalle pätee $F \in \mathrm{MDA}(H_\xi)$, ja sovelletaan semiparametrisia menetelmiä muotoparametrin ξ estimoimiseksi. Olkoon tarkasteltava otos $X_1, \ldots, X_n$, ja vastaavat järjestystunnusluvut nyt $X_{n,n} \leq \cdots \leq X_{1,n}$. Käytetään seuraavia tunnettuja estimaattoreita:

- Hill-estimaattori häntäindeksille $\alpha = \xi^{-1} > 0$:

$$\hat{\alpha}^{(H)} = \hat{\alpha}_{k,n}^{(H)} = \left(\frac{1}{k} \sum_{j=1}^{k} (\ln X_{j,n} - \ln X_{k,n}) \right)^{-1} = \frac{1}{\hat{\xi}^{(H)}}, \quad 2 \leq k \leq n.$$

- Pickands-estimaattori muotoparametrille $\xi \in \mathbb{R}$:

$$\hat{\xi}^{(P)} = \hat{\xi}_{k,n}^{(P)} = \frac{1}{\ln 2} \ln \frac{X_{k,n} - \ln X_{2k,n}}{X_{2k,n} - \ln X_{4k,n}}, \quad 1 \leq k < k_{max} = n/4.$$

- Dekkers–Einmahl–de Haan-estimaattori (DEdH) muotoparametrille $\xi \in \mathbb{R}$:

$$\hat{\xi}^{(D)} = \hat{\xi}_{k,n}^{(D)} = 1 + H_n^{(1)} + \frac{1}{2} \left(\frac{\left(H_n^{(1)}\right)^2}{H_n^{(2)}} - 1 \right)^{-1}, \quad 1 \leq k < n,$$

missä

$$H_n^{(1)} = \frac{1}{k} \sum_{j=1}^{k} (\ln X_{j,n} - \ln X_{k+1,n}),$$

ja

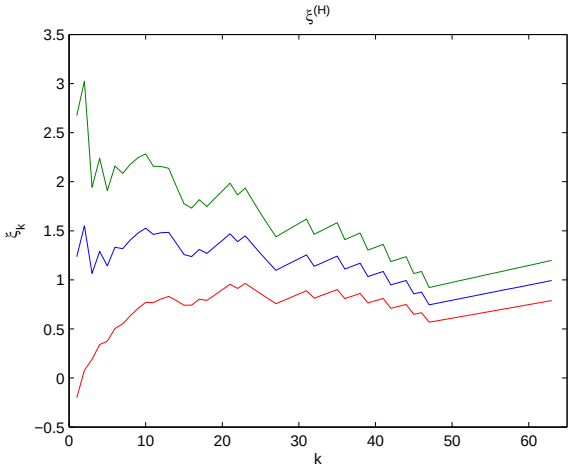
$$H_n^{(2)} = \frac{1}{k} \sum_{j=1}^{k} (\ln X_{j,n} - \ln X_{k+1,n})^2.$$

Estimaattoreilla on tietyin ehdoin hyvät asymptoottiset ominaisuudet (tarkentuvuus, asymptoottinen normaalisuus); estimaattoreista ja niiden ominaisuuksista tarkemmin, ks. [2, luku 6.4].

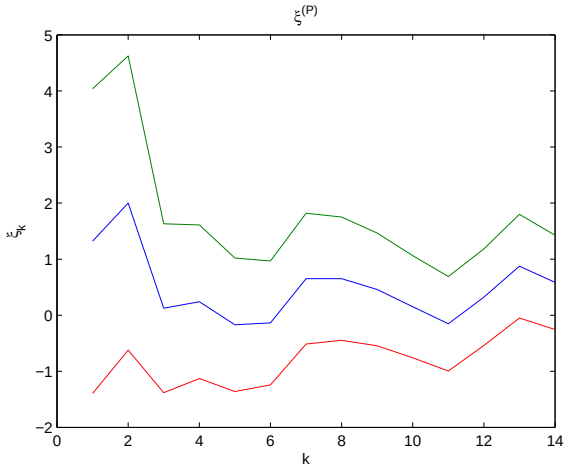
Käytännössä estimaattorien arvot $\hat{\xi}_{k,n}$ piirretään useilla k:n arvoilla ($k \in I$), ja saadusta kuvaajasta $\{(k, \hat{\xi}_{k,n}) : k \in I\}$ etsitään aluetta, josta lähtien estimaatit pysyvät likimain vakaina. Kuvissa 3.3, 3.4 ja 3.5 on esitetty estimaattien $\hat{\xi}^{(H)}$, $\hat{\xi}^{(P)}$ ja $\hat{\xi}^{(D)}$ arvot k:n funktiona 95 %:n luottamusväleineen. Hill-estimaattorin perusteella muotoparametrin arvo vaikuttaisi olevan arviolta välillä [0.8, 1.0], mikä viittaa hyvin paksuhäntäiseen jakaumaan: muistetaan, että kun $\xi > 1/2$, jakauman varianssi ei ole määritelty, ja kun $\xi > 1$, jakauman ensimmäinen momentti (odotusarvo) ei ole määritelty. Nyt siis Hill-estimaattori viittaa vahvasti siihen, että varianssi on ääretön, ja ollaan jopa lähellä tilannetta, että odotusarvo olisi ääretön. Pickands-estimaattorille jää käyttöön niin vähän havaintoja, että tuloksena olevassa kuvaajassa on paljon heiluntaa, mutta sekin näyttäisi viittaavan arvoon $\xi > 1/2$. DedH-estimaattori antaa arvoksi noin 0.6, luottamusvälin ollessa likimain [0.4, 0.8].

3.2.1 Ylitemenetelmä

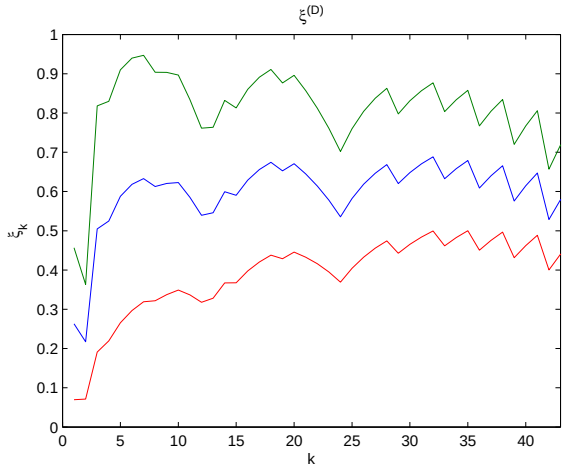
Siirrytään parametriseen menetelmään yleistetyn Pareto-jakauman sovittamisen muodossa. Aloitetaan tarkastelu kynnyksen valinnasta. Kuvaan 3.6 on piirretty



Kuva 3.3: Hill-estimaattori muotoparametrille ξ.



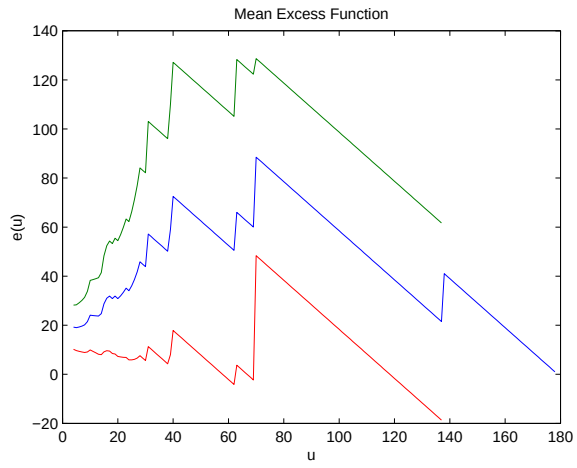
Kuva 3.4: Pickands-estimaattori muotoparametrille ξ.



Kuva 3.5: DEdH-estimaattori muotoparametrille ξ.

ylitteen otoskeskiarvokuvaaja havaintoaineistolle 95 % luottamusväleineen. Kuvasta nähdään, että kuvaaja kasvaa arvoon 30 saakka, jonka jälkeen se etenee hyppäyksin. Kuvaajan käytöksestä suurilla kynnyksen arvoilla, noin arvosta $u = 70$ lähtien, ei voida vetää enää johtopäätöksiä ylitteiden lukumäärän lähestyessä nollaa ja ylitteen odotusarvofunktion arvon mennessä automaattisesti nollaan suurinta havaintoa vastaavan kynnyksen kohdalla. Luottamusvälit huomioiden kynnykseksi u valitaan kuvaajan perusteella arvo 30:n ja 40:n väliltä.

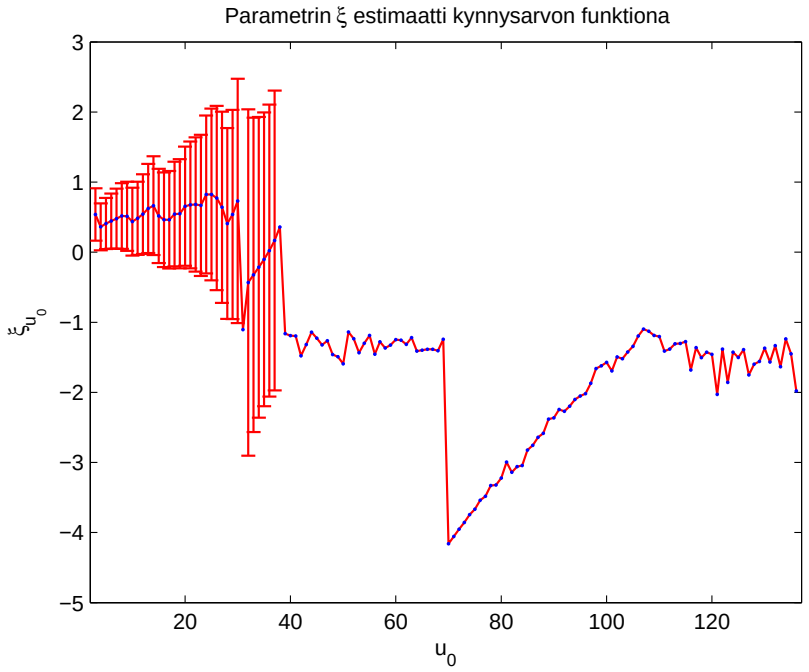
Edellä nähtiin, että onnettomuuskuolemien jakauma on hyvin paksuhäntäinen. Tässä yhteydessä on hyvä muistaa, että mikäli muotoparametrin todelliselle arvolle pätee $\xi > 1$, ei GP-jakauman ensimmäistä momenttia ole olemassa. Tällaisessa tapauksessa datasta voidaan tietenkin yhä laskea havaintojen *otoskeskiarvo*, mutta kuvaajan tulkinta ei ole selvä.



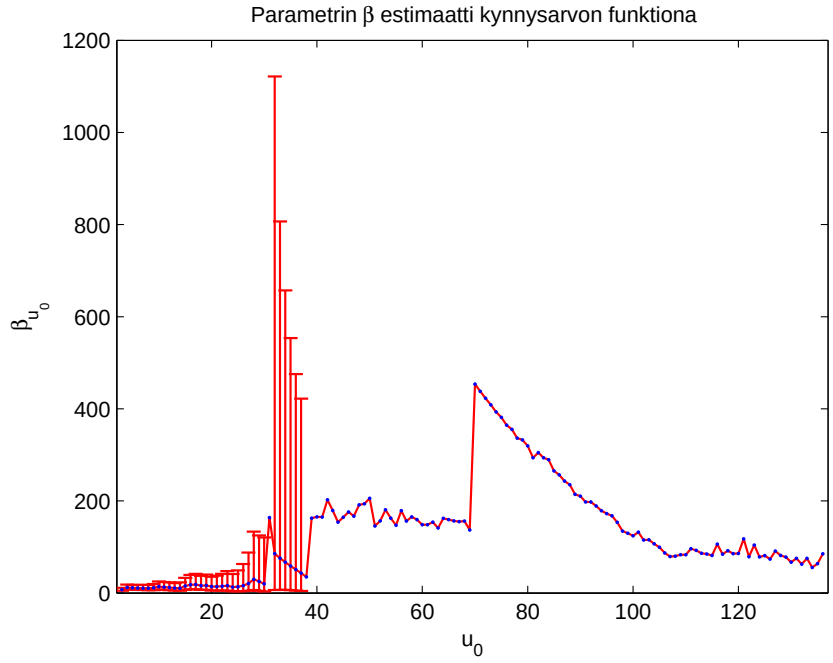
Kuva 3.6: Ylitteen otoskeskiarvokuvaaja onnettomuuskuolemadatalle.

Tarkastellaan seuraavaksi parametriestimaattien stabiilisuutta eri kynnysarvoilla. Kuvissa 3.7, 3.8 ja 3.9 on esitetty parametrien ξ, β ja $\beta^* = \beta - \xi u$ estimaatit eri kynnysarvoilla u sovitetuista GP-jakaumista. Parametrien estimoinnissa kohdattiin numeerisia ongelmia korkeilla kynnyksen arvoilla (eli jakaumaa muutamiin havaintoihin sovitettaessa), mutta kuviin on esimerkin ja havainnollistuksen vuoksi silti piirretty piste-estimaatit laajalta kynnysarvoväliltä. Luottamusvälien laskenta sujui ongelmitta arvoon $u = 30$ asti, ja tuotti jotakin mahdollisen oloista likimain arvoon $u = 40$ asti, mutta tämän jälkeen asymptoottiseen kovarianssimatriisiin perustuvia luottamusvälejä ei käytännössä saatu määritettyä suurimman uskottavuuden menetelmän päätyessä määrittelyalueen rajalla oleviin arvoihin parametriestimaateille.

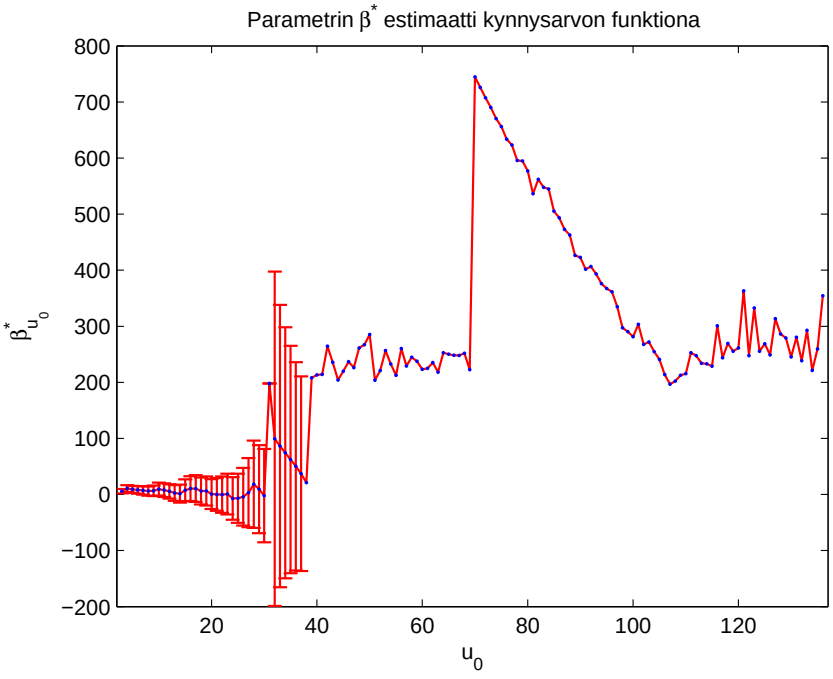
Erityisesti muotoparametrin piste-estimaatti kääntyy vahvasti negatiiviseksi (alle -1) korkeammilla kynnyksen arvoilla. Tämä viittaa suoraan ongelmiin parametrien estimoinnissa. Kuvissa havaitut "hyppyjä" seuraavat likimain lineaarisesti kasvavat tai vähenevät alueet vastaavat datan välejä, joihin ei osu yhtään havaintoa. Parametrin β estimaatin pitäisi olla lineaarinen kynnyksen u suhteen, mikäli $\xi \neq 0$ (kuten tässä tapauksessa voidaan sanoa olevan); ottaen huomioon edellä tyhjistä alueista sanotun, estimaatti $\hat{\beta}$ näyttäisi tosiaankin muuttuvan lähes lineaarisesti arvon $u = 30$ jälkeen. Vastaavasti $\hat{\beta}^*$:n estimaatin tulisi pysyä vakiona alueella, jolla GP-jakauma on hyväksyttävä approksimaatio ylitteiden jakaumalle. Nyt $\hat{\beta}^*$ näyttää pysyvän likimain vakiona arvosta $u = 30$ lähtien. Myös ξ:n kuvaajassa havaittiin tason muutos tämän arvon kohdalla, vaikka estimaatit $\hat{\xi}$ eivät olekaan järkeviä suuremmilla u.



Kuva 3.7: Parametrin ξ estimaatti onnettomuuskuolemadataan sovitetulle GP-jakaumalle kynnysarvon funktiona.

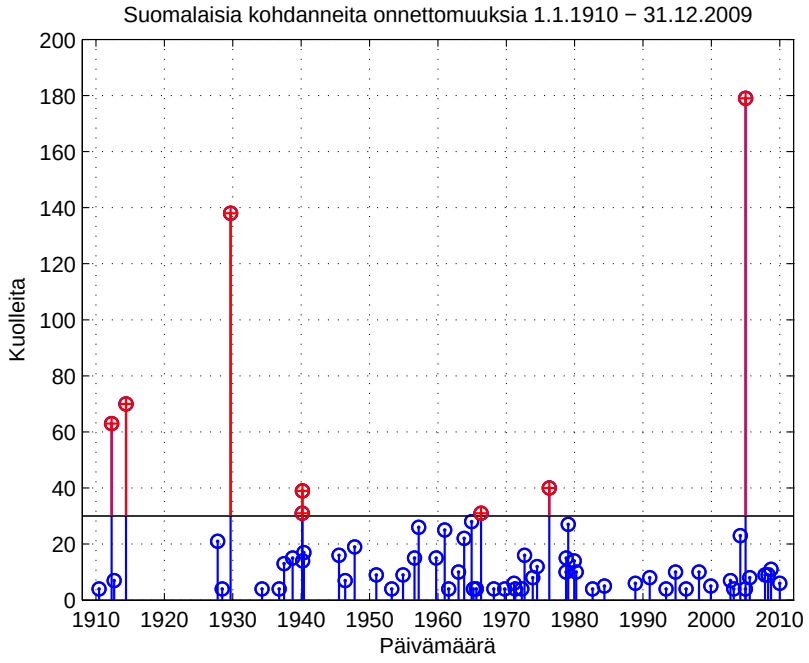


Kuva 3.8: Parametrin β estimaatti onnettomuuskuolemadataan sovitetulle GP-jakaumalle kynnysarvon funktiona.



Kuva 3.9: Parametrin β^* estimaatti onnettomuuskuolemadataan sovitetulle GP-jakaumalle kynnysarvon funktiona.

Esitettyjen tarkastelujen perusteella valitaan kynnystasoksi $u = 30$. Tämän tason ylitteitä on datassa 8 kpl (kuva 3.10), mikä on erittäin vähän tilastollisen analyysin pohjaksi. Vertaillaan saatuja tuloksia sen vuoksi myös koko havaintoaineistoon sovitetun malliin antamiin tuloksiin, jossa siis kynnyksenä on $u = 3$, vaikka tämä kynnys liian pieneltä GP-jakauma-approksimaation perusteltavuutta ajatellen vaikuttaakin.

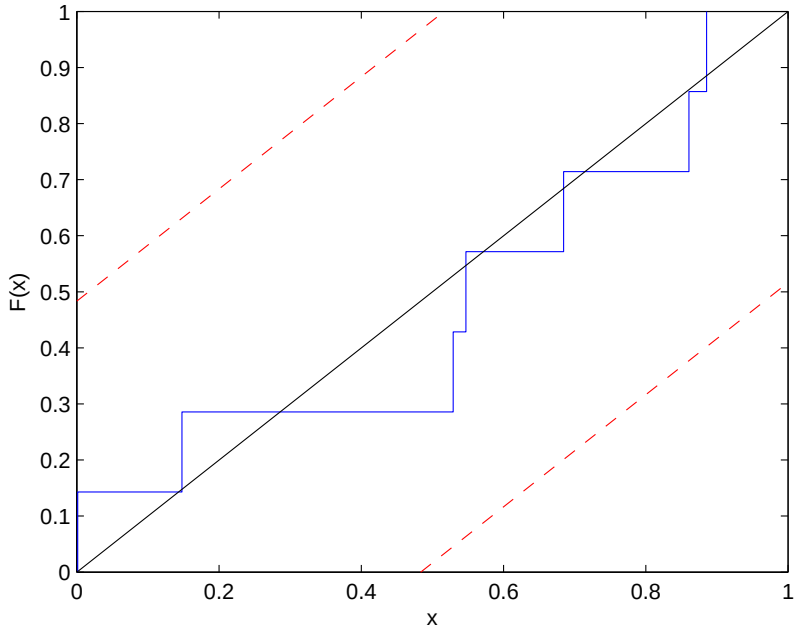


Kuva 3.10: Onnettomuuskuolemadata ja tason $u = 30$ ylitteet.

Edellä tarkasteltiin GP-jakaumaa tason u ylitteiden ehdollisen jakauman (ylitejakauman) mallina. Edellisistä luvuista muistetaan, että ylitemenetelmän taustalla on implisiittisesti oletus, että (korkean tason) ylitteiden lukumäärä tarkasteluväleillä on Poisson-jakautunut. Mikäli ylitysajat eivät noudata Poisson-prosessia, viittaa tämä siihen, ettei ylitedataa voi pitää iid:nä (vaikka pelkät ylitteiden suuruudet erillään aikaulottuvuudesta tarkasteltuna iid olisivatkin). Tällöin iid datalle formuloidut perusmallit eivät automaattisesti sovellu, vaan datan generoineen prosessin kuvaamiseksi saattaa olla tarpeen tarkastella yleisempiä malleja.

Testataan Poisson-jakautuneisuutta alaosion 2.6.2.2 mukaisesti. Mikäli ylitysajat noudattavat Poisson-prosessia, ovat peräkkäisten ylitysaikojen välit eli odotusajat eksponenttijakautuneita. Kuviin 3.11 ja 3.12 on piirretty muunnettuun odotusaikaan perustuvan suureen U_k empiirinen kertymäfunktio tasoja $u = 30$ ja $u = 3$ vastaten. Poisson-oletuksen pätiessä muunnokset U_k noudattavat tasajakaumaa (välillä $[0, 1)$), eli kuvaajan pisteiden tulisi osua lähelle yksikködiagonaalia; punaiset katkoviivat kuvissa ilmoittavat 95 %:n luottamusvälit. Kuvan 3.11 luottamusvälien laajuudesta nähdään, että näin pienellä otoskoolla testillä ei ole voimaa hylätä tasajakautuneisuushypoteesia. Tällä ei kuiten-

kaan tässä tilanteessa ole käytännön merkitystä, sillä muunnettujen suureiden empiirinen jakauma vastaa varsin hyvin tasajakaumaa, eikä evidenssiä ylitysten Poisson-jakautuneisuutta vastaan ilmene. Tason $u = 3$ ylitteitä tarkastellessa ei myöskään ole perusteita hylätä ylitysten lukumäärien Poisson-jakaumaoletusta.

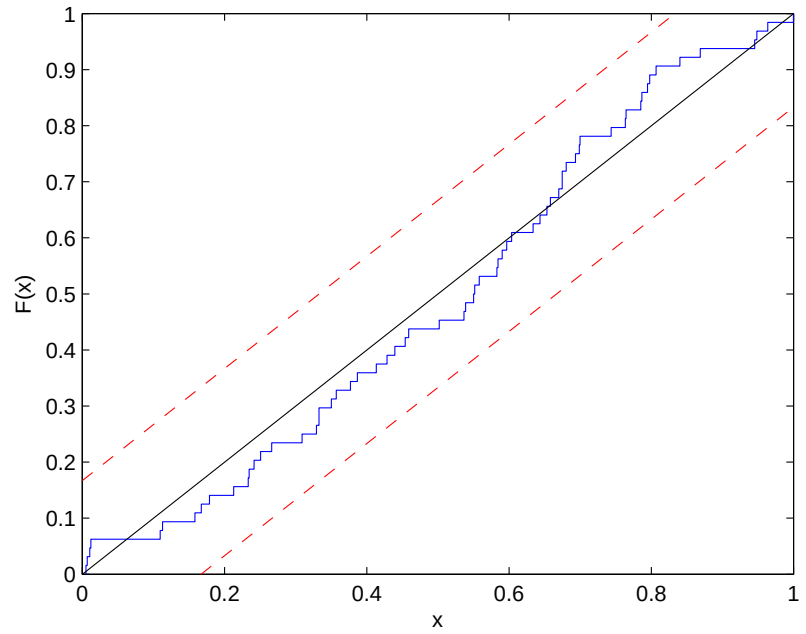


Kuva 3.11: Muunnettuihin odotusaikoihin perustuvan suureen U_k empiirinen jakauma (sininen viiva) tasajakaumaa vastaan luottamusväleineen; $u = 30$.

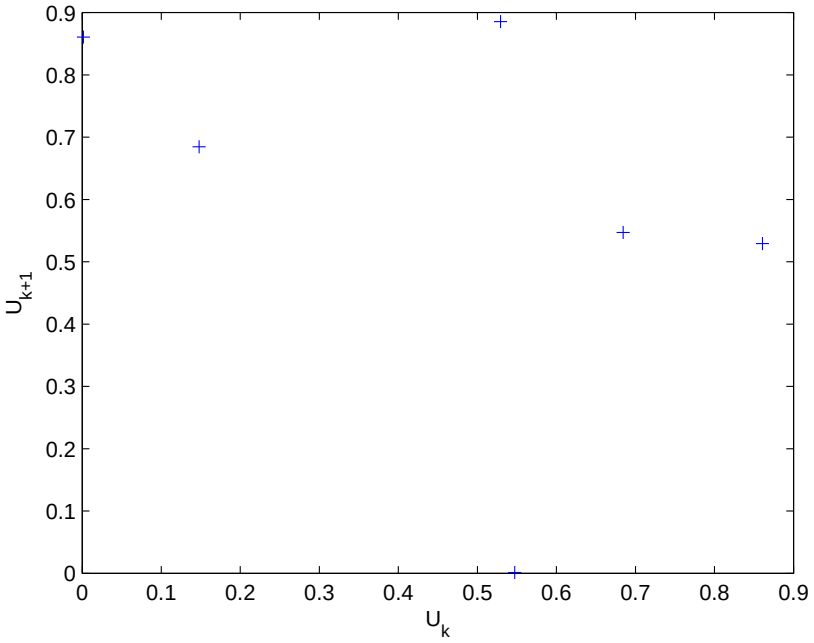
Testataan seuraavaksi (vierekkäisten) ylitysaikojen riippumattomuutta muunnettuihin odotusaikoihin U_k perustuen. Mikäli viereiset odotusajat ovat riippumattomia – ja siten myös ylitysajat ovat – tulisi pisteiden (U_k, U_{k+1}) olla tasan jakautuneita yksikköneliössä $[0, 1) \times [0, 1)$, eli pisteiden tulisi jakautua satunnaisesti alueelle. Kuviin 3.13 ja 3.14 on piirretty pisteet tasoja $u = 30$ ja $u = 3$ vastaten. Ensimmäisen kuvan perusteella ei voida sanoa paljonkaan havaintojen vähyyden vuoksi; perusteita epäillä riippumattomuutta ei toisaalta myöskään ilmene. Jälkimmäinen kuva, vastaten tasoa $u = 3$, näyttää varsin satunnaiselta. Tämän ja edellisen tarkastelun johtopäätös siis on, että ylityksien lukumäärää voidaan pitää Poisson-jakautuneena (jolloin ylitykset siis sattuvat homogeenisen Poisson-prosessin mukaisesti).

Jatketaan ylitemallin estimointiin suurimman uskottavuuden menetelmällä. Taulukossa 3.2 on esitetty saadut parametriestimaatit kynnystasoille $u = 30$ ja $u = 3$.

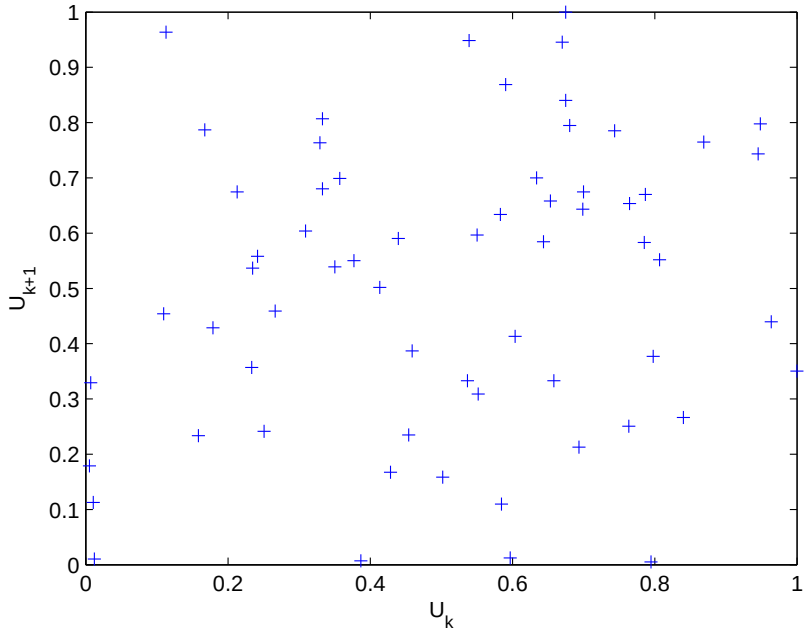
Muotoparametrin ξ estimaattien arvot ovat samaa suuruusluokkaa kuin edellä semiparametristen menetelmien perusteella arvioitiin. Nähdään, että piste-estimaattien $\hat{\xi}$ arvot ovat molemmilla kynnystasoilla yli 0.5, mikä viittaa siihen,



Kuva 3.12: Muunnettuihin odotusaikoihin perustuvan suureen U_k empiirinen jakauma (sininen viiva) tasajakaumaa vastaan luottamusväleineen; $u = 3$.



Kuva 3.13: U_k vs. U_{k+1}; $u = 30$.



Kuva 3.14: U_k vs. U_{k+1}; $u = 3$.

Taulukko 3.2: GP-jakauman parametriestimaatit asymptoottiseen keskivirheeseen perustuvine luottamusväleineen.

Kynnys	$\boldsymbol{u = 30}$		$\boldsymbol{u = 3}$	
Parametri	SUE	95% luottamusväli	SUE	95% luottamusväli
ξ	0.731	$[-1.01, 2.47]$	0.538	$[0.163, 0.912]$
β	19.8	$[3.26, 121]$	7.45	$[4.86, 11.4]$

ettei jakaumien toinen momentti ole olemassa. Luottamusvälit ovat kuitenkin laajat, erityisesti korkeamman kynnyksen $u = 30$ tapauksessa; tässä havaintoja on niin vähän (8 kpl) ja estimaattorin otosvarianssi siten niin suuri, että saadut luottamusvälit eivät kerro juuri mitään. Alarajan negatiivisuus johtuu jälleen asymptoottiseen keskivirheeseen perustuvien luottamusvälien symmetrisyydestä parametrille ξ.

Tarkastellaan vertailun vuoksi profiiliuskottavuusmenetelmällä saatavia luottamusvälejä muotoparametrille. Nämä on esitetty taulukossa 3.3 alla. Nähdään, että luottamuväli ei ole enää symmetrinen: kynnystasolla $u = 3$ alaraja on suurempi kuin keskivirheeseen perustuen, samoin yläraja. Tämä on tyypillistä vertailtaessa näitä kahta menetelmää luottamusvälien rakentamiseen tapauksessa $\xi \geq 0$: pakosta symmetrinen keskivirheeseen perustuva menetelmä antaa usein epäuskottavan pienen alarajan, ja toisaalta profiiliuskottavuuden käyttö huomioi ylärajaan usein liittyvän suuren epävarmuuden paremmin. Kynnyksen $u = 30$ tapauksessa luottamusvälin ylärajaestimaatti on huomattavasti suurempi kuin edellä saatu, ja alarajan määritys numeerisesti epäonnistuu. Tässä tapauksessa menetelmän käyttö ei siis anna lisäinformaatiota.

Taulukko 3.3: Muotoparametrin ξ luottamusvälit profiiliuskottavuuteen perustuen.

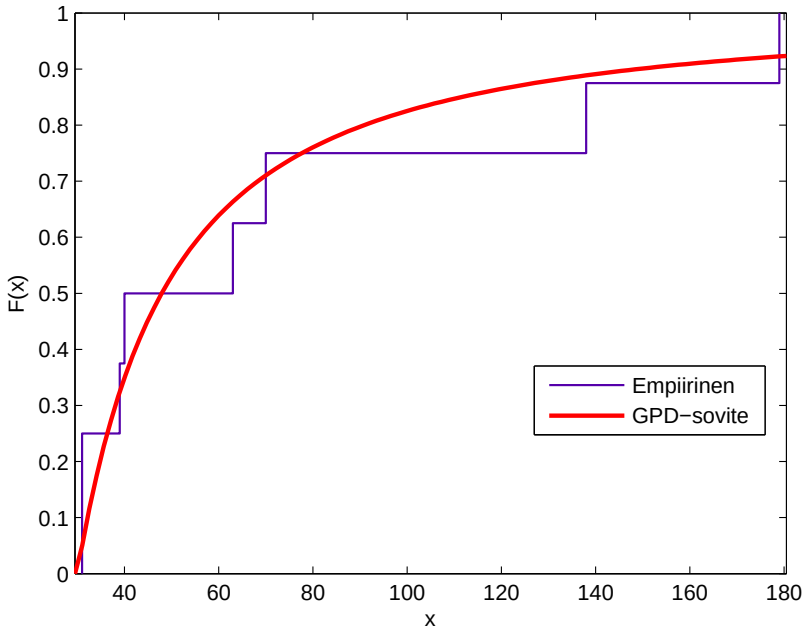
Kynnys, u	**$\hat{\xi}$**	**95%:n luottamusväli**
3	0.538	$[0.238, 1.01]$
30	0.731	$[-, 3.85]$

Kuvissa 3.15 ja 3.16 on vertailtu sovitettua GP-jakaumaa havaintoihin perustuvaan empiiriseen jakaumaan näiden kertymäfunktioita tarkastelemalla. Kuvan 3.15 sovite, vastaten valittua tasoa $u = 30$, näyttää sopivan havaintoihin varsin hyvin, huolimatta siitä että empiirinen kertymäfunktio on hyvin sahalaitainen havaintojen vähyydestä johtuen. Kuvan 3.16 koko dataan ($u = 3$) sovitettu GP-jakauma näyttää myös vastaavan hyvin havaittua.

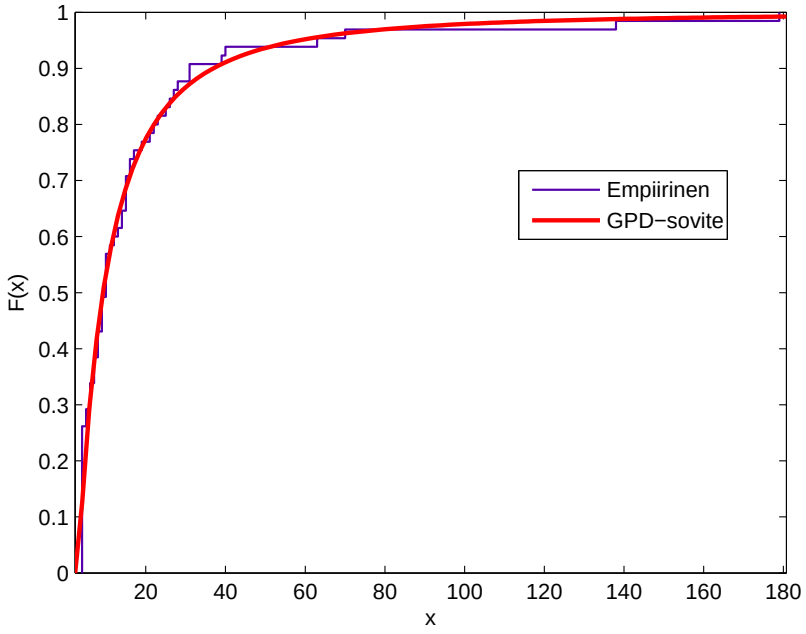
Jatketaan ylitemallin sopivuuden tarkastelua tutkimalla tuttuun tapaan todennäköisyys- ja kvantiilikuvaajia. Kuviin 3.17 ja 3.18 on piirretty kuvaajat tason $u = 30$ ylitteisiin sovitetulle GP-jakaumalle, ja kuvissa 3.19 ja 3.20 on vastaavasti todennäköisyys- ja kvantiilikuvaaja tason $u = 3$ ylitteisiin sovitetulle mallille.

Korkeampaa tasoa vastaavan mallin todennäköisyyskuvaaja (kuva 3.17) näyttää sopivan kohtalaisen hyvin vähiin häntähavaintoihin. Matalan tason tapauksessa sopivuus jakauman alkupäässä on heikohko, mutta näyttää paranevan jakauman oikeaa häntää kohti mentäessä (kuva 3.19); erilaisesta skaalauksesta johtuen hännän sopivuus ei kuitenkaan tule selvästi esiin kuvasta, toisin kuin edellä. Matalan tason kuvaajan voidaan katsoa antavan viitteitä siitä, ettei asymptoottiseen argumenttiin perustuva GP-jakaumamalli ole perusteltu näin alhaisella kynnystasolla.

Kvantiilikuvaajista paljastuu selvemmin mallien sopivuus ylitteisiin. Havaitaan, että tason $u = 30$ ylitteisiin sovitetulla mallillakaan sopivuus äärimmäisiin havaintoihin ei näytä erityisen hyvältä. Matalan tason $u = 3$ kohdalla tämä on vielä hieman heikompi. Tämän epäyhteensopivuuden aiheuttaa olennaisesti kak-

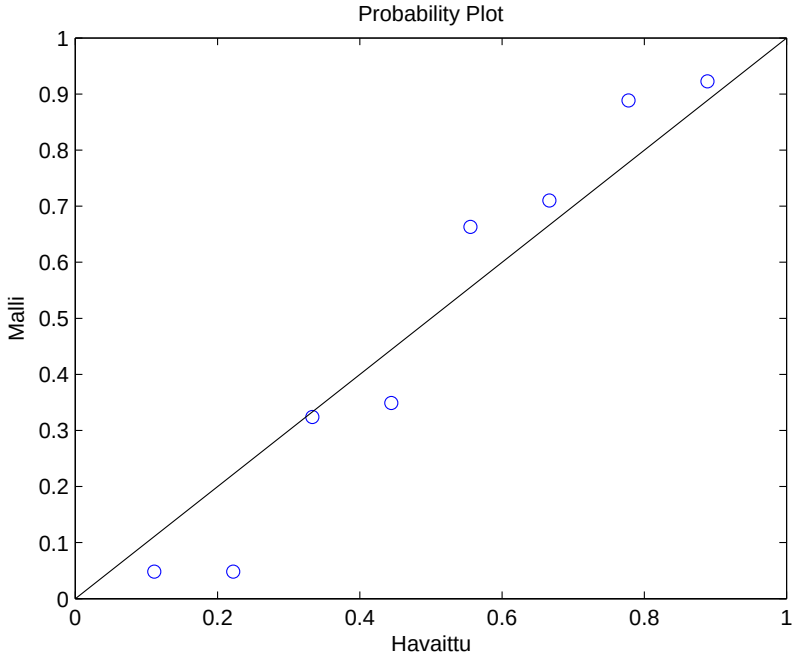


Kuva 3.15: Onnettomuuskuolemien tason $u = 30$ henkeä ylittävien havaintojen empiirinen kertymäfunktio vs. GPD-sovite.



Kuva 3.16: Onnettomuuskuolemien tason $u = 3$ henkeä ylittävien havaintojen empiirinen kertymäfunktio vs. GPD-sovite.

si äärimmäistä havaintoa, jotka ovat kooltaan omassa luokassaan muihin havaintoihin verrattuna (ks. kuva 3.10). Tarkastellussa aineistossa nämä todella ovat äärimmäisiä havaintoja suhteessa muuhun dataan.

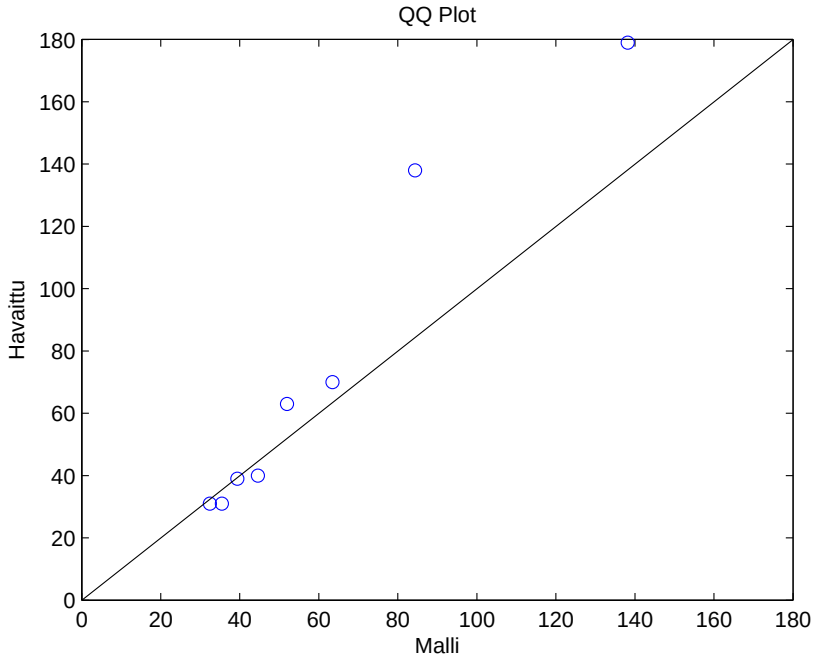


Kuva 3.17: Todennäköisyyskuvaaja onnettomuuskuolemadataan sovitetulle GP-mallille; $u = 30$.

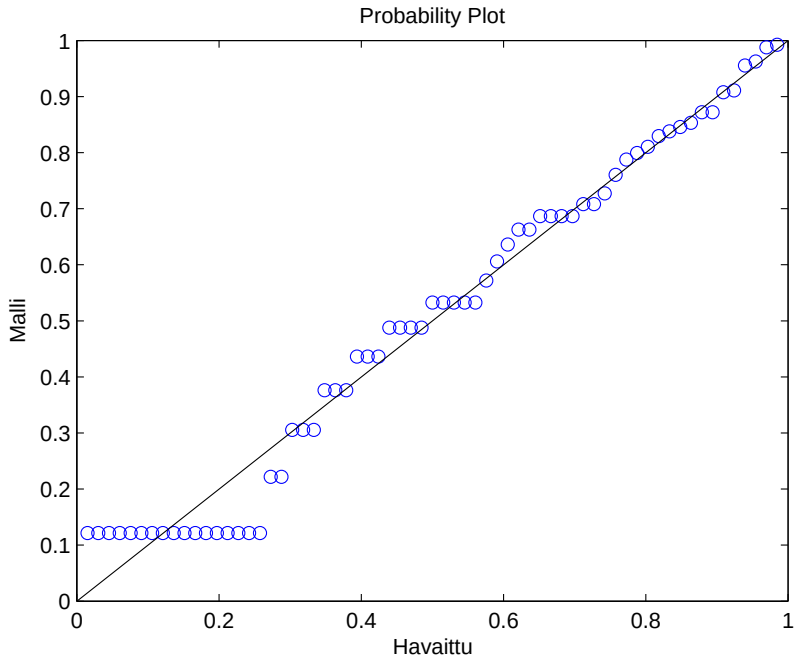
Kuvissa 3.21 ja 3.22 on esitetty onnettomuuskuolemien toistumistasokuvaajat tasojen $u = 30$ ja $u = 3$ ylitteisiin sovitetuille malleille vuositasolla. Kuviin on piirretty myös havaittu data. Sopivuus havaintoihin etenkin korkeamman kynnyksen tapauksessa vaikuttaa kohtuulliselta, ja erityisesti havainnot ovat 95 %:n asymptoottisten delta-menetelmään perustuvien luottamusvälien sisällä. Myös matalamman kynnyksen kohdalla likipitäen kaikki havainnot sisältyvät saatuun 95 %:n luottamusväliin.

Yllä olevista toistumistasokuvaajista havaitaan mm., että alempi 95 %:n luottamusväli menee luottamusvälien symmetrisyydestä johtuen toistumisperiodin kasvaessa negatiiviseksi, mikä ei tietenkään ole mahdollista taustalle olevaa ilmiötä ajatellen. Tarkastellaan vielä profiiliuskottavuusmenetelmään perustuvia luottamusvälejä, joilla päästään yleensä parempaan tarkkuuteen. Tilan säästämiseksi esitetään kuvaajat vain korkeampaa kynnystasoa vastaavalle mallille: kuvassa 3.23 on profiiliuskottavuus 100-vuoden toistumistasolle ja kuvassa 3.23 vastaavasti 1 000-vuoden toistumistasolle. Molemmista kuvaajista havaitaan luottamusvälien erittäin vahva epäsymmetrisyys, mikä johtuu jakauman oikeaa häntää koskevasta suuresta epävarmuudesta.

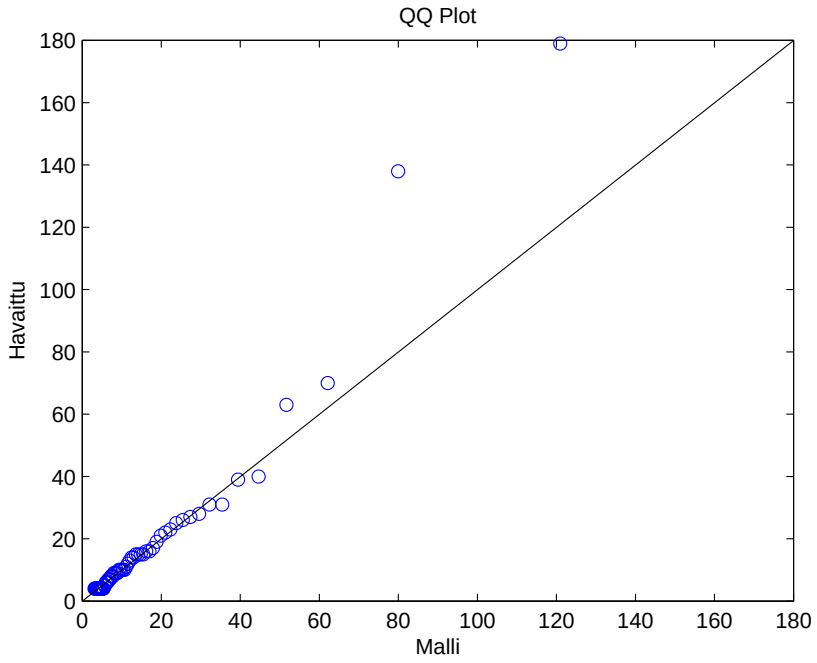
Taulukossa 3.4 on vertailtu delta- ja profiiliuskottavuusmenetelmillä saatuja luottamusvälejä eri malleissa. Luottamusvälejä ja taulukkoon kerättyjä piste-



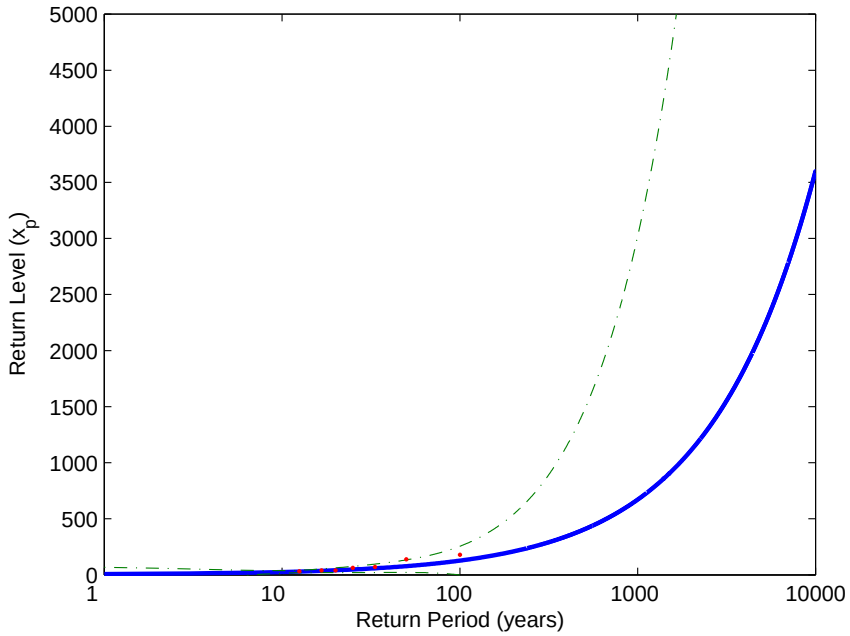
Kuva 3.18: Kvantiilikuvaaja onnettomuuskuolemadataan sovitetulle GP-mallille; $u = 30$.



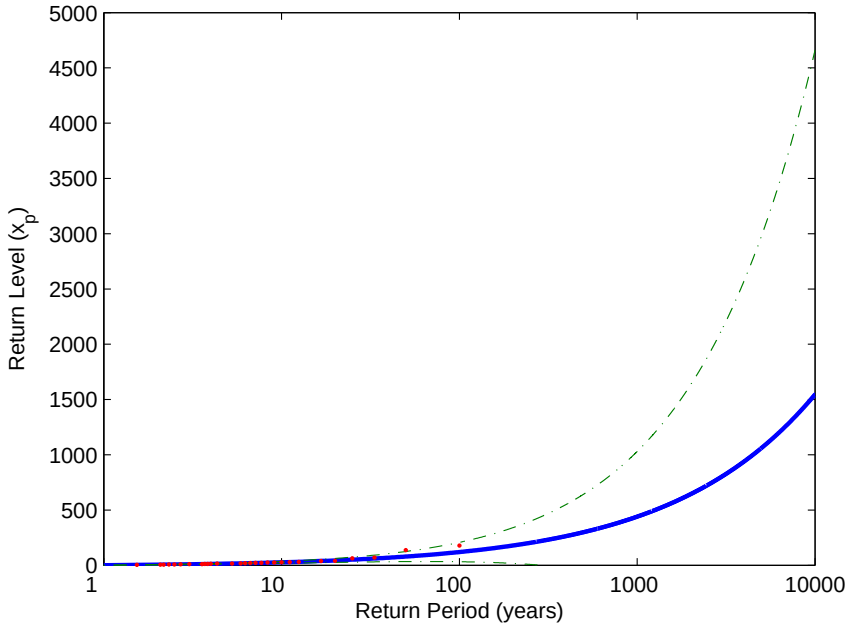
Kuva 3.19: Todennäköisyyskuvaaja onnettomuuskuolemadataan sovitetulle GP-mallille; $u = 3$.



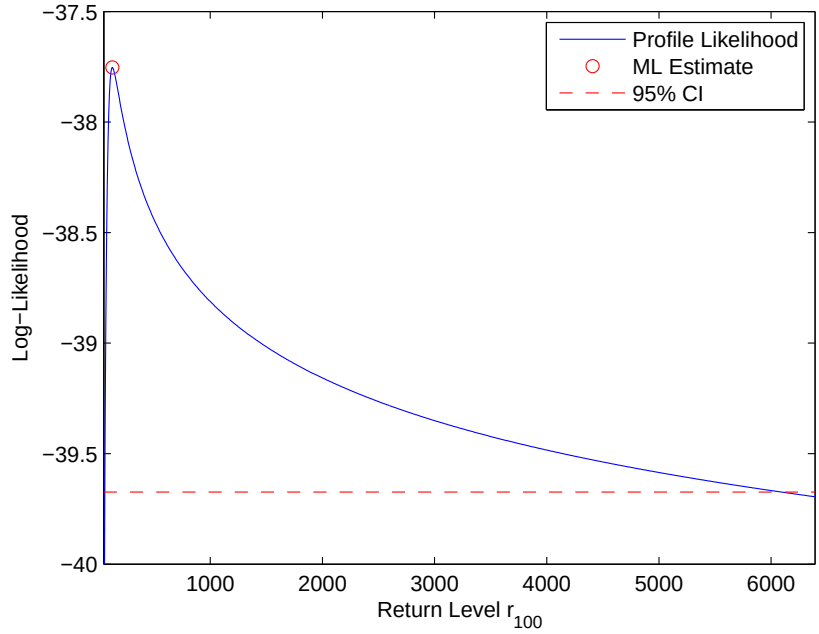
Kuva 3.20: Kvantiilikuvaaja onnettomuuskuolemadataan sovitetulle GP-mallille; $u = 3$.



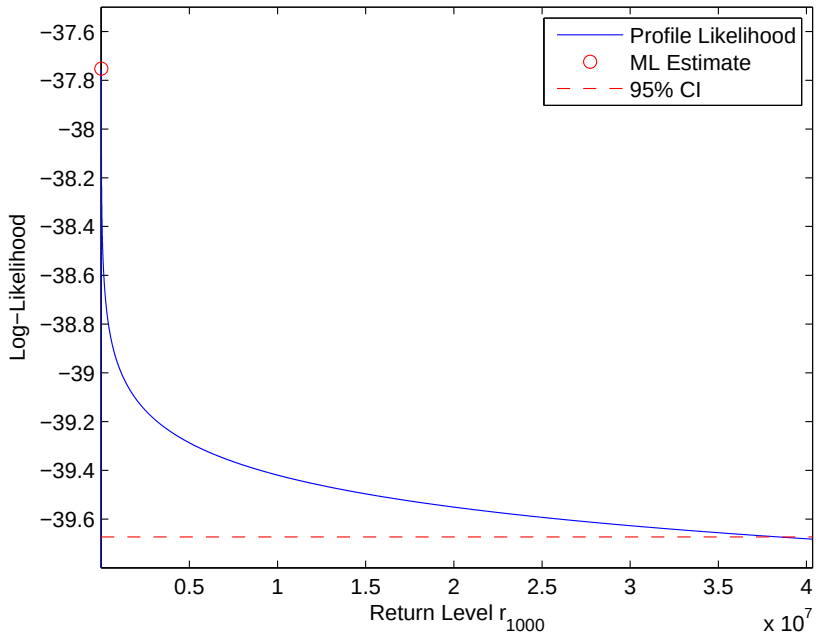
Kuva 3.21: Onnettomuuskuolemien toistumistasokuvaaja GPD-mallissa ($u = 30$) luottamusväleineen.



Kuva 3.22: Onnettomuuskuolemien toistumistasokuvaaja GPD-mallissa ($u = 3$) luottamusväleineen.



Kuva 3.23: Profiiliuskottavuus onnettomuuskuolemien 100-vuoden toistumistasolle GPD-mallissa ($u = 30$).



Kuva 3.24: Profiiliuskottavuus onnettomuuskuolemien 1 000-vuoden toistumistasolle GPD-mallissa ($u = 30$).

estimaatteja tarkastellessa täytyy noudattaa varovaisuutta suurempia toistumisperiodeja tulkittaessa, sillä ekstrapolointiin havaintojen ulkopuolelle liittyy tässä sovelluksessa erittäin suurta epävarmuutta, kuten edellä useaan otteeseen havaittiin. Havainnollistuksen vuoksi niinkin suuri toistumisperiodi kuin 10 000 vuotta on kuitenkin otettu mukaan taulukkoon: On selvää, että tällaisesta tasosta ei pystytä 100 vuoden havaintodatan perusteella sanomaan realistisesti ottaen juuri mitään puhtaasti tilastollisen mallin keinoin, mutta luvut yhdessä muiden toistumistasojen kanssa antavat kuitenkin tuntumaa malliin.

Korkeampaa kynnystasoa $u = 30$ vastaavan mallin kohdalla 1 000- ja 10 000-vuoden toistumistasojen luottamusvälit eroavat valtavasti delta- ja profiiliuskottavuusmenetelmien välillä. Profiiliuskottavuusmenetelmän antamat luottamusvälin alarajat ovat sinänsä järkeviä, vaikka vaikuttavatkin aivan liian pieniltä kun huomioon otetaan – datan ilmentämän informaation ulkopuolinen – tietämys taustalla olevan ilmiön luonteesta eli fysikaalisesti mahdollisten katastrofien suuruudesta. 1 000- ja 10 000-vuoden kohdalla profiiliuskottavuusmenetelmän luottamusvälien asymmetrisyys on jo äärimmäisen suurta, ja täytyy tulkita niin, että dataan sovitetun GP-mallin perusteella käytännössä minkä tahansa suuruiset onnettomuudet ovat mahdollisia.

Vaikka kvantiilikuvaaja korkeamman tason $u = 30$ ylitteisiin sovitetulle mallille ei näytäkään erityisen hyvältä kahden äärimmäisen havainnon osalta, voidaan mallia silti pitää kohtuullisena kaikki muut tarkastelut huomioiden. Havaittu data ei yksinkertaisesti sisällä riittävästi informaatiota sen kokoluokan onnetto-

Taulukko 3.4: Delta-menetelmään ja profiliuskottavuuteen perustuvat 95 %:n luottamusvälit onnettomuuskuolemien toistumistasoille GPD-malleissa.

$u = 3$	Delta-menetelmä		Profiiliuskottavuus	
Toistumisperiodi	SUE	95% CI	SUE	95% CI
10	27	$[18, 36]$	27	$[19, 42]$
100	120	$[32, 210]$	120	$[63, 410]$
1 000	440	$[-150, 1030]$	440	$[142, 4130]$
10 000	1540	$[-1570, 4660]$	1540	$[270, 41800]$
$u = 30$	Delta-menetelmä		Profiiliuskottavuus	
Toistumisperiodi	SUE	95% CI	SUE	95% CI
10	26	$[14, 38]$	26	-
100	127	$[-1, 260]$	127	$[60, 6090]$
1 000	670	$[-1680, 3030]$	670	$[140, 38 \cdot 10^6]$
10 000	3600	$[-20600, 27800]$	3600	$[180, 270 \cdot 10^9]$

muuksista (suhteessa muihin havaintoihin) joita tarkastelujakson aikana on kuitenkin havaittu kaksi. Tarkastellaan seuraavaksi mahdollista keinoa ongelman lieventämiseksi.

3.2.2 Laajennettu aineisto

Kuten edellä havaittiin, suomalaisia kohdanneista suuronnettomuuskuolemista on niin vähän havaintoja, että – *havaintoihin perustuvien* – tilastollisten menetelmien käyttö on ongelmallista. Ääriarvoteoriaan perustuvia menetelmiä soveltaen katastrofikuolemille saatiin rakennettua todennäköisyysjakauma, joka sinänsä vaikuttaa järkevältä käytettävissä olevien kriteerien suhteen tarkasteltuna, mutta johon liittyy suuri epävarmuus laajojen luottamusvälien muodossa.

Kuvatunlainen tilanne on tuttu jälleenvakuutuksesta, jossa saattaa syntyä tilanne, että tietyntyyppinen riski pitäisi hinnoitella, vaikkei siihen liittyvistä tappioista ole kokemusta (experience) olemassa. Tällöin standardi menettelytapa on etsiä markkinoilta tietoa mahdollisimman samankaltaisista riskeistä, ja käyttää tätä informaatiota tukena hinnoittelussa.

Pyritään seuraavassa tarkentamaan mallia käyttämällä laajennettua tilastoaineistoa, joka sisältää Ruotsin kokemuksen. Yhdistetään siis edellä kuvattuun suomalaisia koskevaan aineistoon vastaava, ruotsalaisia kohdanneista (ja yli kolme henkeä vaatineista) onnettomuuskuolemista kerätty aineisto samalta aikaväliltä.[8] Yhdistämisen katsotaan olevan perusteltua, koska Ruotsia voidaan pitää lähinnä Suomea muistuttavana maana (alueena, populaationa), ja näiden kahden maan maantieteelliset ja taloudelliset olosuhteet ovat toisiinsa verrattavia (vrt. myös kohdan 3.1 yleiseen keskusteluun).

Koska tarkoituksena kuitenkin on rakentaa todennäköisyyspohjainen malli kuvaamaan suomalaisten katastrofikuolemien sattumista eikä suomalaisten ja ruot-

[8] Aineisto koostuu julkisista lähteistä kerätyistä tiedoista; ks. ruotsinkielinen Wikipedia (sv.wikipedia.org), kategoria "Olyckor", artikkeli "Lista över katastrofer efter antalet döda svenskar".

salaisten yhteistä katastrofikuolleisuutta, vaikuttaa ilmeiseltä että jonkinlaista skaalausta tarvitaan ennen yhdistetyn aineiston tai sen perusteella estimoidun mallin käyttöä. Koska onnettomuuskuolemia maiden välillä voidaan pitää toisistaan riippumattomina, estimoidaan vahinkojen suuruusjakauma suoraan yhdistetystä aineistosta (ks. kuitenkin seuraava kappale). Sen sijaan yhdistetystä aineistosta estimoitu vahinkojen sattumisfrekvenssi on selvästi liian suuri, koska se sisältää molempien maiden onnettomuudet tarkastelujaksolta. Vaihtoehtoina on käyttää Suomen aineiston perusteella määritettyä sattumisfrekvenssiä, tai esimerkiksi maiden "keskiarvoista" frekvenssiä. Tähän palataan tarkemmin tuonnempana.

Aineistoja yhdistäessä täytyy huomioida muutamia seikkoja. Erityisesti yhteiset vahingot voi olla tarpeen siivota pois aineistosta niin, ettei sama tapahtuma esiinny datassa useaan kertaan. Vahingojen suuruuksien (lasketaanko vahingot yhteen, otetaanko mukaan suurempi vahinko, jne.) sekä vahinkotapahtumien poiston sattumisfrekvenssivaikutuksen tarkka käsittely riippuu sovelluksesta ja mallin käyttötarkoituksesta. Tässä tapauksessa aineistossa oli kaksi päällekäistä onnettomuutta, matkustajalaiva Estonian uppoaminen 28.9.1994 (suomalaisia kuoli 10, ruotsalaisia 552) ja Intian valtameren tsunami 26.12.2004 (suomalaisia kuoli 179, ruotsalaisia 543). Otetaan Estonian osalta mukaan suoraan Ruotsin aineiston kuolemien määrä (ja poistetaan vastaava Suomen aineiston rivi yhdistetystä aineistosta), koska on ajateltavissa, että tämän kokoluokan onnettomuus voisi tapahtua myös suomalaisia täynnä olevalle laivalle. Tsunamikuolemien osalta pätee sama periaate; koska Ruotsin väestö on kuitenkin n. 1.75-kertainen Suomen väestöön verrattuna, myös ruotsalaisia turisteja voidaan ajatella olevan enemmän. Skaalataan siis tsunamissa kuolleiden ruotsalaisten lukumäärä väkilukujen suhteella (vuoden 2011 lopussa), jolloin saadaan kuolintapausten lukumääräksi kokonaisluvuksi pyöristettynä 316. Käytetään tätä lukua yhdistetyssä aineistossa. Mitä havaintojen poistamisen frekvenssivaikutukseen tulee, vaikutusta ei ole, jos käytetään vain Suomen aineistosta estimoitua frekvenssiä. Sen sijaan poistetut havainnot tulee huomioida (lisätä takaisin vahinkojen lukumäärään), jos käytetään maiden keskiarvofrekvenssiä.

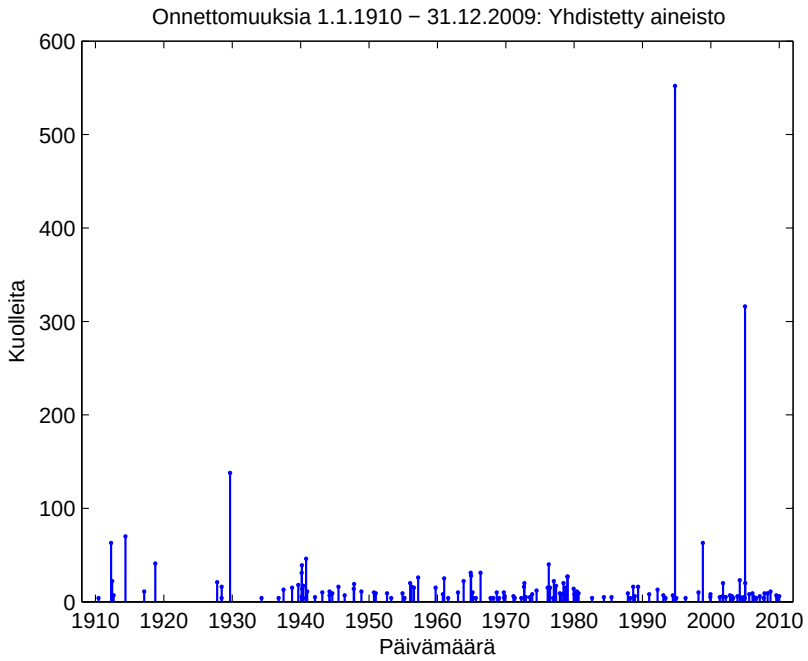
Taulukossa 3.5 on esitetty joitakin tunnuslukuja edellä kuvatulla tavalla saadulle yhdistetylle onnettomuuskuolema-aineistolle. Havaintoja tarkastelujaksolta 1.1.1910 – 31.12.2009 on yhteensä 139 kpl, ja suurin näistä on 552 kuollutta Estonian uppoamisessa.

Taulukko 3.5: Tilastollisia tunnuslukuja suomalaisten ja ruotsalaisten yhdistetylle onnettomuuskuolemadatalle.

n	min	max	mediaani	moodi	keskiarvo	keskihajonta	IQR
139	4	552	9	4	19.6	54.5	11.0

Kuvassa 3.25 on visualisoitu havaintoaineisto. Nähdään, että tässä tapauksessa suurin havainto eroaa selkeästi muista, ja kaksi suurinta taas jäljelle jäävistä. Myös kolmanneksi suurin havainto eroaa suuruusluokaltaan lopuista. Jo kuvan perusteella nähdään, että alla oleva datan generoineen ilmiön todennäköisyysjakauma vaikuttaa erittäin paksuhäntäiseltä.

Tarkastelut kynnystason valitsemiseksi (ei näytetty) viittaavat yhdistetyn ai-



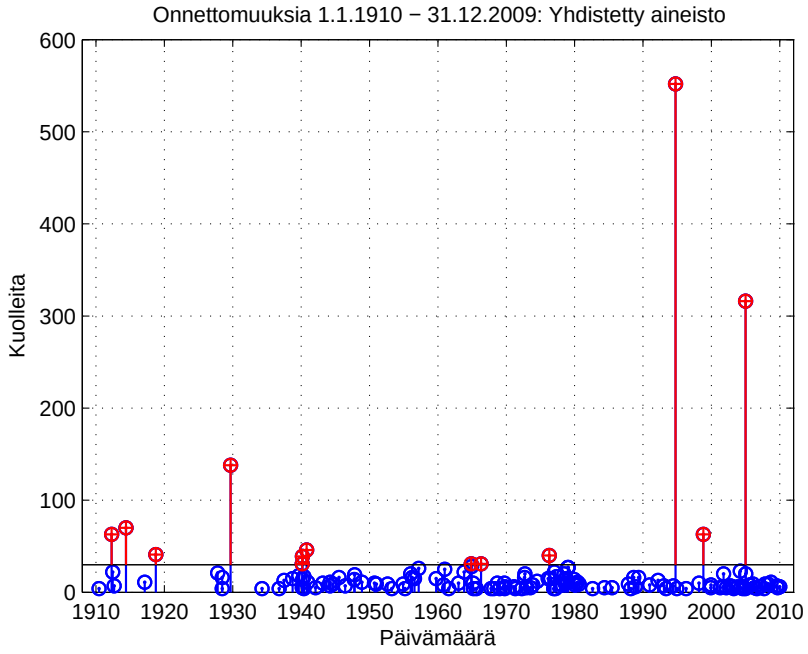
Kuva 3.25: Suuronnettomuuksia päivämäärän mukaan, yhdistetty aineisto.

neistonkin tapauksessa siihen, että taso $u = 30$ voisi olla perusteltu. Valitaan tämä.

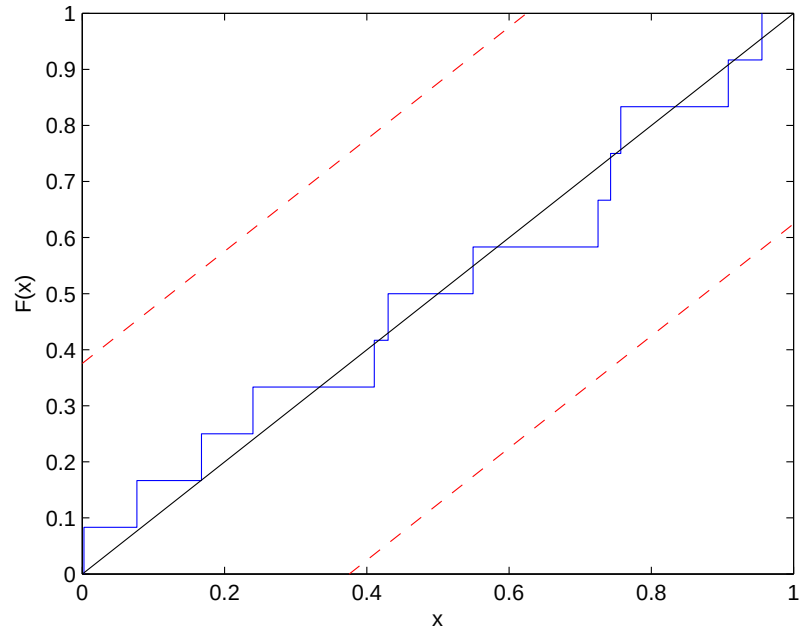
Aineistossa on 13 tason 30 ylittävää onnettomuutta (ks. kuva 3.26). Tämä on suhteellisesti suuri lisäys verrattuna pelkkään suomalaisia koskevaan onnettomuuskuolema-aineistoon, jossa oli 8 saman tason ylitystä, mutta absoluuttisesti edelleen hyvin pieni havaintojen lukumäärä tilastolliseen analyysiin. Kuten edellisessä osiossa, tarkastellaan vertailun vuoksi myös koko aineistoon sovitettua mallia, joka siis vastaa kynnystasoa $u = 3$.

Tarkastellaan ylitteiden sattumisten Poisson-jakautuneisuutta edellisen osion mukaisesti havaintodataa koskevan iid-oletuksen testaamiseksi. Kuvissa 3.27 ja 3.28 on esitetty muunnettuihin ylitysten välisiin odotusaikoihin perustuvan suureen empiirinen jakauma. Kuviin pätevät pitkälti samat huomiot kuin edellisessä osiossa tarkastellun datan kohdalla. Tason $u = 30$ tapauksessa ylitykset vaikuttavat selvästi Poisson-jakautuneilta. Tason $u = 3$ kohdalla puolestaan on havaittavissa, ettei evidenssi ylitysten Poisson-jakautuneisuudesta (muunnettujen odotusaikojen tasajakautuneisuudesta) ole tässä tapauksessa aivan yhtä vahvaa, vaikka empiirinen jakauma 95 % luottamusvälien sisällä pysyykin, ja siten Poisson-jakautuneisuusoletusta ei hylätä tarkastellulla merkitsevyystasolla.

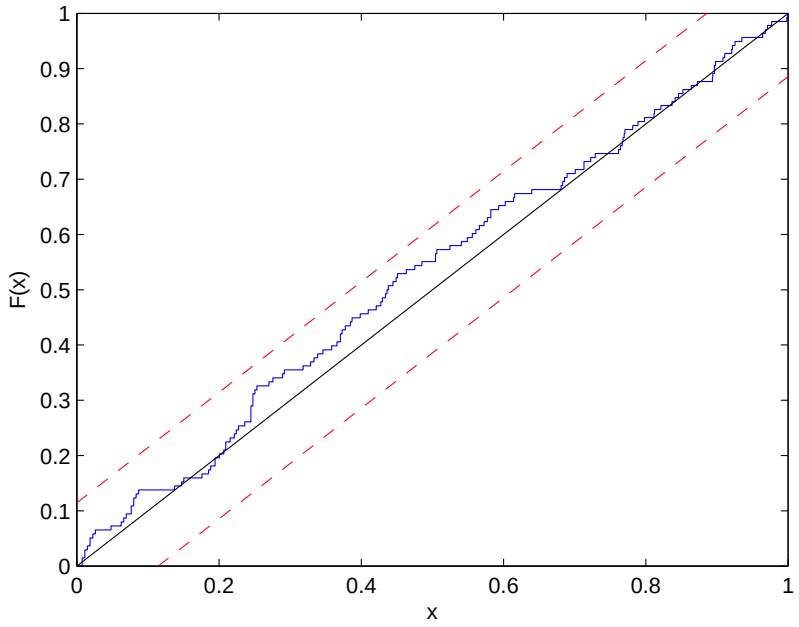
Vierekkäisten ylitysaikojen riippumattomuuden tarkastelu edellisen osion tapaan (ei näytetty) johtaa samaan lopputulokseen kuin pelkästään suomalaisia koskevan datan kohdalla; tason $u = 30$ ylitteitä on niin vähän, että tätä vastaava kuvaaja ei ole kovin informatiivinen, mutta kumpaakaan tasoa vastaava



Kuva 3.26: Yhdistetty onnettomuuskuolemadata ja tason $u = 30$ ylitteet.



Kuva 3.27: Muunnettuihin odotusaikoihin perustuvan suureen U_k empiirinen jakauma (sininen viiva) tasajakaumaa vastaan luottamusväleineen; $u = 30$.



Kuva 3.28: Muunnettuihin odotusaikoihin perustuvan suureen U_k empiirinen jakauma (sininen viiva) tasajakaumaa vastaan luottamusväleineen; $u = 3$.

kuvaaja ei anna aihetta hylätä ylitysten Poisson-jakautuneisuusoletusta tämän testin perusteella. Johtopäätös on siis – kuten edellisessä osiossa – että ylitykset sattuvat homogeenisen Poisson-prosessin mukaisesti. Tämä tarkoittaa, että ylitemenetelmän (ja POT-mallin) taustaoletukset täyttyvät näiltä osin, ja tason u ylityksien sattumista sekä toisaalta ylitteiden suuruutta voidaan mallintaa erikseen.

Sovitetaan yleistetty Pareto-jakauma tasojen $u = 30$ ja $u = 3$ ylitteisiin. Taulukossa 3.6 on esitetty suurimman uskottavuuden menetelmällä saadut GP-jakauman parametriestimaatit, sekä näiden keskivirheeseen perustuvat 95 % luottamusvälit. Nähdään, että matalan tason $u = 3$ ylitteisiin sovitettu malli antaa lähes saman estimaatin muotoparametrille ξ kuin edellisessä osiossa saatiin, mutta lyhyemmillä luottamusväleillä. Tason $u = 30$ ylitteisiin sovitettu malli antaa muotoparametrin estimaatiksi $\hat{\xi} = 1.26$, mikä viittaa äärimmäisen paksuhäntäiseen jakaumaan; kun $\xi > 1$, ei jakauman ensimmäinen momenttikaan ole olemassa.

Taulukko 3.6: GP-jakauman parametriestimaatit asymptoottiseen keskivirheeseen perustuvine luottamusväleineen.

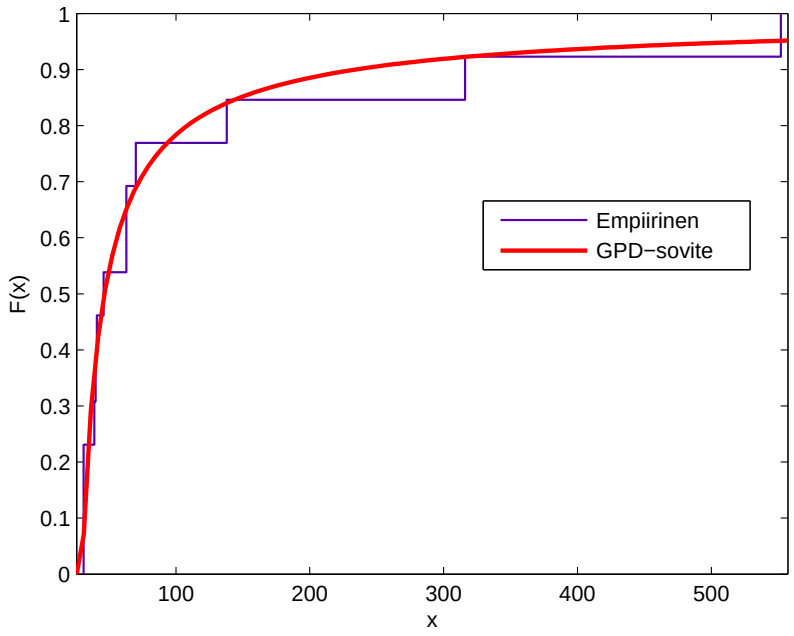
Kynnys	$u = 30$		$u = 3$	
Parametri	SUE	95% luottamusväli	SUE	95% luottamusväli
ξ	1.26	$[-0.086, 2.60]$	0.522	$[0.294, 0.750]$
β	15.1	$[4.20, 54.0]$	6.83	$[5.22, 8.93]$

Profiiliuskottavuusmenetelmään perustuvat luottamusvälit muotoparametrille on esitetty taulukossa 3.7 alla. Matalammalle tasolle luottamusväli on hyvin samankaltainen kuin keskivirheeseen perustuva. Korkeamman tason kohdalla luottamusväli sen sijaan on selvästi epäsymmetrisempi, ja erityisesti alaraja on sinällään uskottava, toisin kuin taulukossa 3.6. 95 % luottamusvälin alarajat ovat profiiliuskottavuusmenetelmällä hyvin lähellä toisiaan molempiin kynnystasoihin sovitettujen mallien kohdalla.

Taulukko 3.7: Muotoparametrin ξ luottamusvälit profiiliuskottavuuteen perustuen.

Kynnys, u	**$\hat{\xi}$**	**95%:n luottamusväli**
3	0.522	$[0.329, 0.790]$
30	1.26	$[0.362, 3.46]$

Kuvaan 3.29 on piirretty tason $u = 30$ ylitteiden empiirinen kertymäfunktio ja tätä vastaava sovitettu GP-jakauma. Empiirisen kertymäfunktion diskreetin luonteen huomioiden sovitus näyttää hyvältä.

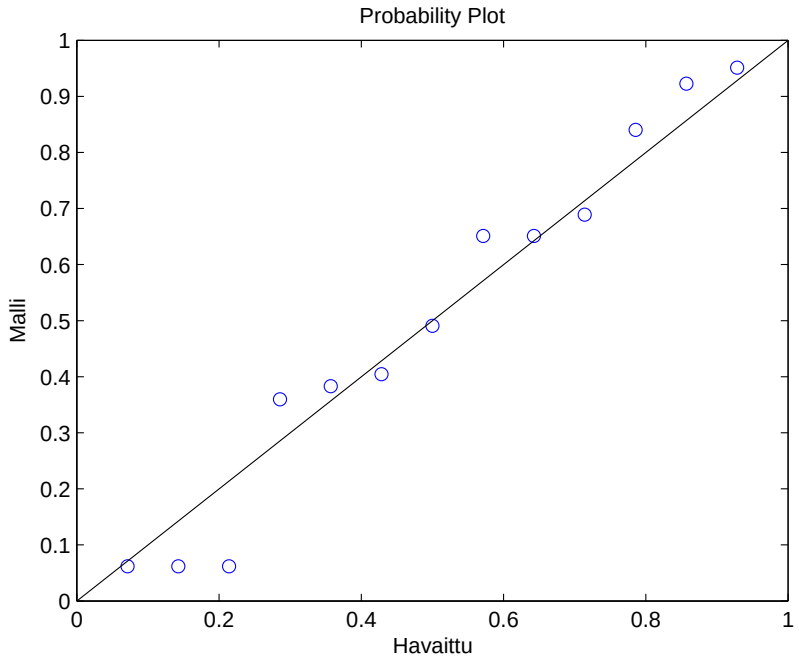


Kuva 3.29: Onnettomuuskuolemien tason $u = 30$ henkeä ylittävien havaintojen empiirinen kertymäfunktio vs. GPD-sovite; yhdistettu aineisto.

Myös todennäköisyyskuvaajan perusteella tason $u = 30$ ylitteitä vastaavan mallin sopivuus havaintoihin on varsin hyvä (kuva 3.30). Kvantiilikuvaaja (kuva 3.31) puolestaan tuo esiin saman ongelman kuin edellisessä osiossa: koko (tason u) ylitedataan sovitettu malli ei sovikaan kovin hyvin kahteen (tai kolmeen) äärimmäisimpään havaintoon. Näiden äärihavaintojen suuruusluokka eroaa, kuten edellä nähtiin, selvästi kaikista muista havainnoista. Tämä viittaa siihen,

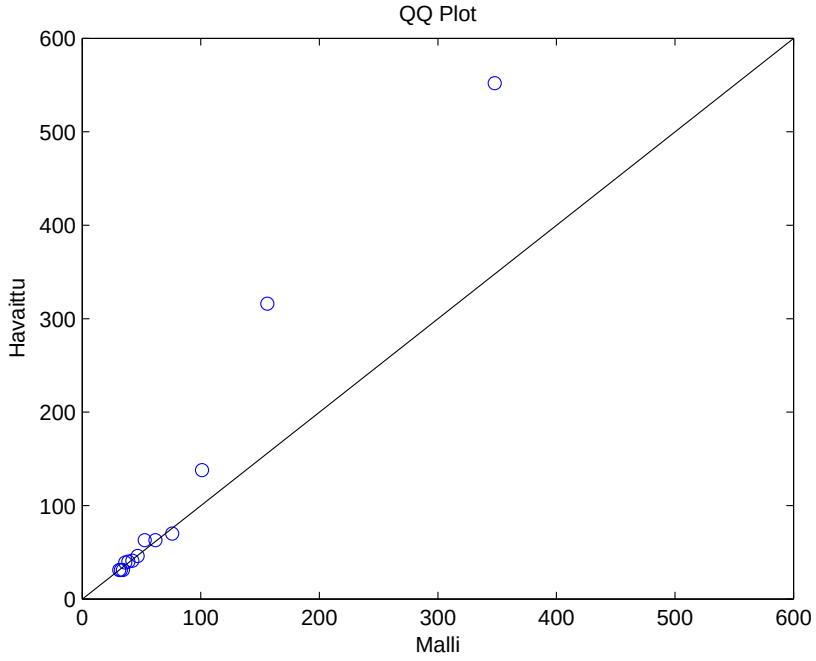
että näiden äärihavaintojen mallintamisen ja GP-jakauman asymptoottisen luonteen näkökulmasta pienimmät GP-jakauman sovitukseen mukaan otetut havainnot ovat mahdollisesti liian pieniä, ja käytettyä kynnystasoa u tulisikin ehkä nostaa. Tätä kokeiltaessa kynnystason kasvattaminen ei kuitenkaan olennaisesti muuttanut tilannetta, ja tietyn pisteen jälkeen GP-jakauman sovittaminen suurimman uskottavuuden menetelmällä ei enää onnistu havaintojen vähyyden vuoksi. Päädytään siis jälleen siihen johtopäätökseen, että – koska äärimmäisiä havaintoja tässä sovelluksessa ei pidetä poikkeavina (outliereinä), vaan kaikista arvokkaimpina havaintoina – käytettävissä olevat havainnot eivät yksinkertaisesti sisällä riittävästi informaatiota ilmiön taustalla olevan todennäköisyysjakauman hännän tarkkaa tilastollista kuvaamista varten.

Vertailun vuoksi kuvissa 3.32 ja 3.33 on esitetty todennäköisyys- ja kvantiilikuvaajat koko aineistoon, eli tason $u = 3$ ylitteisiin, sovitetulle GP-jakaumalle. Kvantiilikuvaaja viittaa selvästi siihen, että taso $u = 3$ on liian matala, jotta asymptoottiseen argumenttiin perustuva GP-jakauma-approksimaatio eli ylitejakauman approksimoiminen GP-jakaumalla olisi perusteltua.

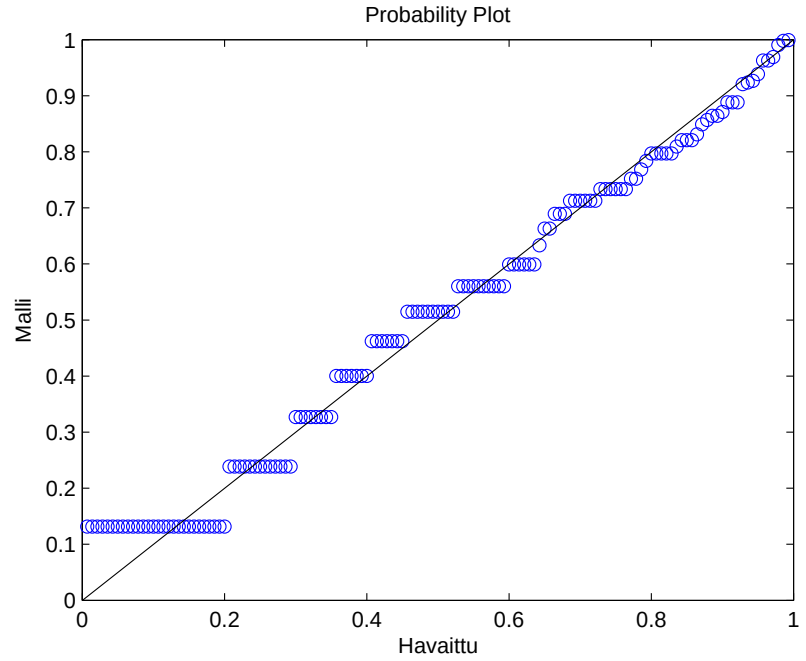


Kuva 3.30: Todennäköisyyskuvaaja yhdistettyyn onnettomuuskuolemadataan sovitetulle GP-mallille; $u = 30$.

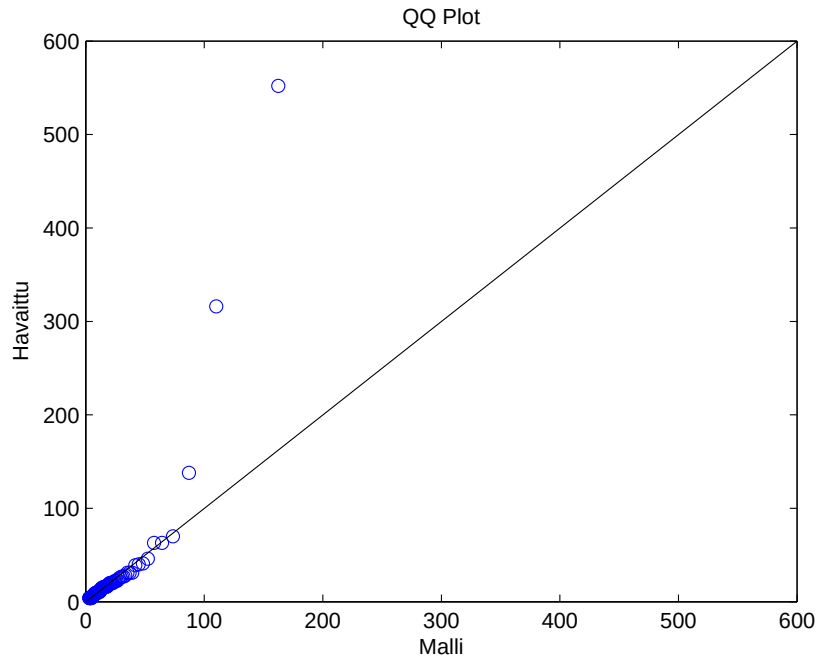
Tasoa $u = 30$ vastaavan mallin perusteella saatu toistumistasokuvaaja on piirretty kuvaan 3.34. Vaikka estimoidut toistumistasot eroavat havaituista korkeimpien havaintojen kohdalla, ovat havainnot delta-menetelmään perustuvien 95 %:n luottamusvälien sisällä lukuun ottamatta toiseksi suurinta havaintoa, joka jää niukasti luottamusvälin ulkopuolelle. Tasoa $u = 3$ vastaavan mallin kohdalla (kuvaajaa ei näytetty) suuremmat havainnot jäävät selkeästi mallin implikoimien toistumistason luottamusvälien ulkopuolelle.



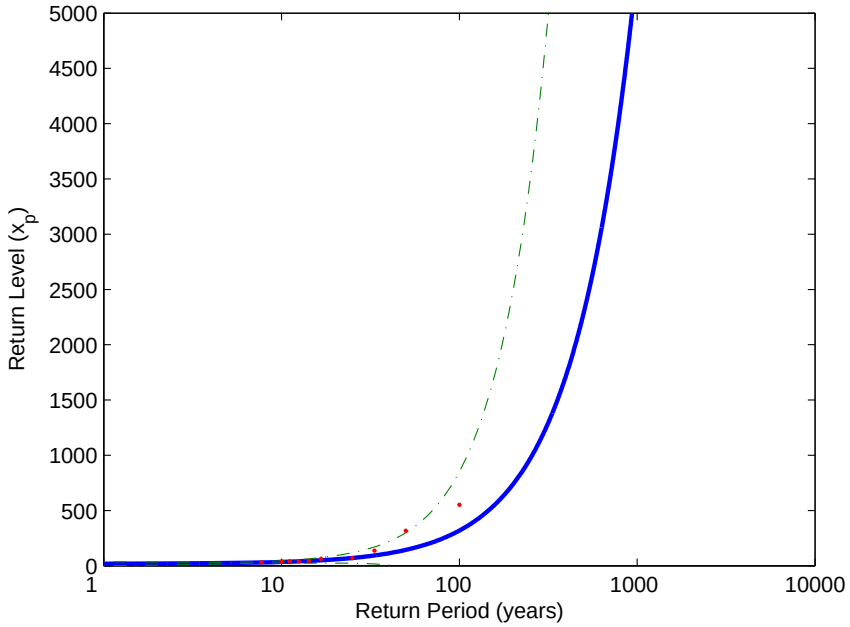
Kuva 3.31: Kvantiilikuvaaja yhdistettyyn onnettomuuskuolemadataan sovitetulle GP-mallille; $u = 30$.



Kuva 3.32: Todennäköisyyskuvaaja yhdistettyyn onnettomuuskuolemadataan sovitetulle GP-mallille; $u = 3$.

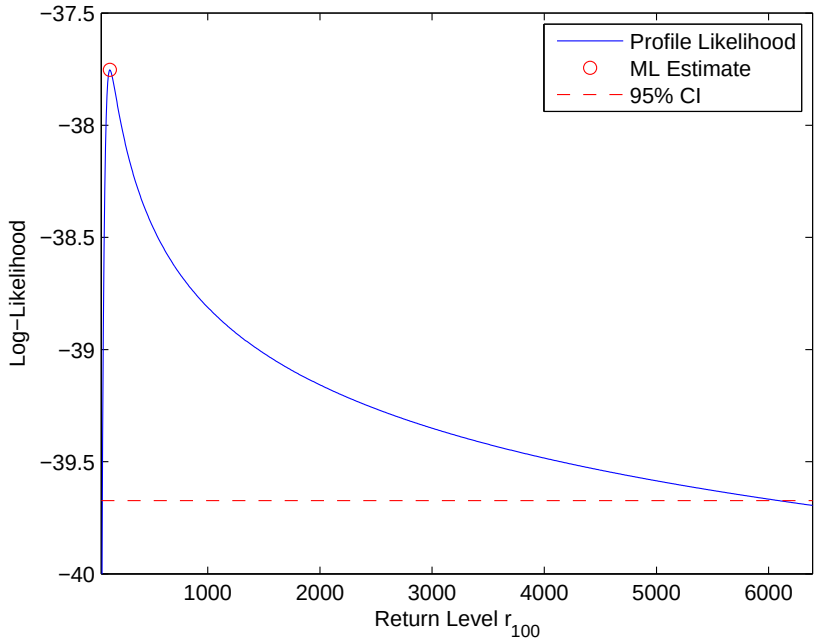


Kuva 3.33: Kvantiilikuvaaja yhdistettyyn onnettomuuskuolemadataan sovitetulle GP-mallille; $u = 3$.



Kuva 3.34: Onnettomuuskuolemien toistumistasokuvaaja GPD-mallissa ($u = 30$) luottamusväleineen; yhdistetty aineisto.

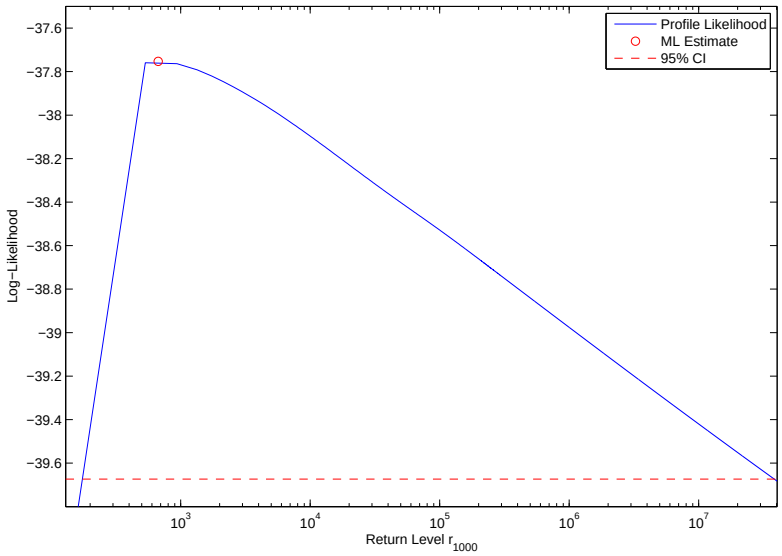
Kuvissa 3.35 ja 3.36 on esitetty profiiliuskottavuuskuvaaja sekä profiiliuskottavuuteen perustuvien luottamusvälien määräytyminen 100- ja 1 000-vuoden toistumistasoille tason $u = 30$ ylitteisiin perustuvassa mallissa. Jo sadan vuoden toistumistasoa vastaavasta kuvaajasta nähdään luottamusvälien erittäin voimakas asymmetrisyys ja ylärajaan liittyvä epävarmuus. Tuhannen vuoden toistumistasoa vastaavan profiiliuskottavuuskuvaajan kohdalla nämä piirteet ovat niin voimakkaita, että havainnollisuuden säilyttämiseksi kuvaaja on välttämätöntä esittää logaritmisella skaalalla.



Kuva 3.35: Profiiliuskottavuus onnettomuuskuolemien 100-vuoden toistumistasolle yhdistettyyn aineistoon perustuvassa GPD-mallissa ($u = 30$).

Taulukossa 3.8 on piste-estimaatit sekä delta- ja profiiliuskottavuusmenetelmiin perustuvat luottamusvälit toistumisperiodeita 10, 100, 1 000 ja 10 000 vuotta vastaaville toistumistasoille. Tason $u = 30$ ylitteisiin sovitetun mallin antamien toistumistasojen lisäksi taulukossa on esitetty vertailun vuoksi myös koko aineistoon sovitetun mallin antamat luvut, vaikka tätä mallia ei voikaan edellä esitetyn tarkastelun perusteella pitää perusteltuna kuvauksena havaintojen jakauman hännästä.

Toistumistasoihin pätee jo edellisen osion lopussa vastaavan tarkastelun yhteydessä sanottu. Korkeampia toistumistasoja koskeviin lukuarvoihin on syytä suhtautua asianmukaisella varauksella, ottaen huomioon ekstrapolaation aste. Profiiliuskottavuusmenetelmällä saatavat luottamusvälit ovat yleensä käytännössä aina tarkempia kuin delta-menetelmällä saadut keskivirheeseen perustuvat, mutta tässä tapauksessa luottamusvälit ovat suuremmilla toistumistasoilla varsin epäinformatiivisia. Olennaisesti mallin mukaan miten suurten tahansa onnettomuuksien sattuminen voi olla mahdollista, tai tällaisten onnettomuuksien



Kuva 3.36: Profiiliuskottavuus onnettomuuskuolemien 1 000-vuoden toistumistasolle yhdistettyyn aineistoon perustuvassa GPD-mallissa ($u = 30$).

Taulukko 3.8: Delta-menetelmään ja profiliuskottavuuteen perustuvat 95 %:n luottamusvälit onnettomuuskuolemien toistumistasoille yhdistettyyn aineistoon perustuvissa GPD-malleissa.

$u = 3$	Delta-menetelmä		Profiiliuskottavuus	
Toistumisperiodi	SUE	95% CI	SUE	95% CI
10	42	$[30, 53]$	42	$[31, 61]$
100	160	$[69, 260]$	160	$[94, 390]$
1 000	560	$[24, 1100]$	560	$[220, 2400]$
10 000	1900	$[-710, 4500]$	1900	$[510, 15000]$
$u = 30$	Delta-menetelmä		Profiiliuskottavuus	
Toistumisperiodi	SUE	95% CI	SUE	95% CI
10	35	$[24, 46]$	35	$[31, 45]$
100	320	$[-220, 860]$	320	$[100, 18000]$
1 000	5400	$[-17000, 28000]$	5400	$[360, 47 \cdot 10^6]$
10 000	98000	$[-560 \cdot 10^3, 750 \cdot 10^3]$	98000	$[1000, 130 \cdot 10^9]$

mahdollisuutta ei voida sulkea käytetyn mallin perusteella pois malliparametreihin liittyvä epävarmuus huomioiden.

Tilanteen ollessa tämä sovitettua GP-mallia voidaan kuitenkin pitää kohtuullisena kuvauksena onnettomuuskuolemadatasta käytettävissä olevan tilastollisen informaation pohjalta.

3.2.2.1 Skaalattu malli

Edellä tarkasteltu malli, mukaan lukien tarkastellut toistumistasot, koski yhdistettyä aineistoa, eli kuvasi (muutamaa pientä muokkausta lukuun ottamatta) Suomen ja Ruotsin yhteistä katastrofikuolleisuutta. Tämän luvun sovelluksen tavoitteena on kuitenkin ensisijaisesti kuvata suomalaisten onnettomuuskuolemia. Kuten edellisen osion alussa mainittiin, tämän saavuttamiseksi onnettomuuksien sattumisfrekvenssiä tulee muuttaa maiden yhteenlasketusta sattumisfrekvenssistä Suomen kokemusta vastaavaksi. Onnettomuuksien suuruutta (tason u yläpuolella) kuvaavaksi jakaumaksi sen sijaan otetaan suoraan yhdistetystä aineistosta estimoitu GP-jakauma.

Todennäköisyys tapahtumalle, että taso u ylitetään, *ehdolla* että taso $u_0 < u$ on ylitetty, voidaan estimoida suoraan havaintodatasta: valitaan $u_0 = 3$ (datan kynnystaso) ja olkoon tasojen u_0 ja u ylityksien lukumäärät n ja k. Tällöin saadaan estimaatti

$$\mathbb{P}(X > u | X > u_0) = \frac{k}{n}.$$

Ylityksien suhteellinen osuus k/n on myös yo. todennäköisyyden suurimman uskottavuuden estimaattori, kun ylityksien lukumäärän oletetaan olevan binomi- tai Poisson-jakautunut.

Sattumistodennäköisyyksien tai –frekvenssien tarkastelua vaikeuttaa se seikka, ettei havaintoaineistossa ole selkeää struktuuria. Havainnot eivät ole tasavälisiä, eikä niihin pystytä liittämään yksikäsitteistä aikayksikköä: vaikka esimerkiksi havainnot on kerätty päivätasolla, eikä aineistossa ole useampaa samana päivänä tapahtunutta onnettomuutta, tämä ei kuitenkaan tarkoita sitä, etteikö onnettomuuksia voisi sattua useampia samana päivänä. Koska suurien onnettomuuksien sattuminen kuitenkin on hyvin harvinaista, valitaan tarkastelun perusaikayksiköksi yksi vuosi, mikä on kätevää myös tuloksien tulkinnan ja esittämisen kannalta. Tällöin vuositason todennäköisyys sille, että datan kynnystaso $u_0 = 3$ ylitetään, voidaan havaintoaineiston perusteella estimoida

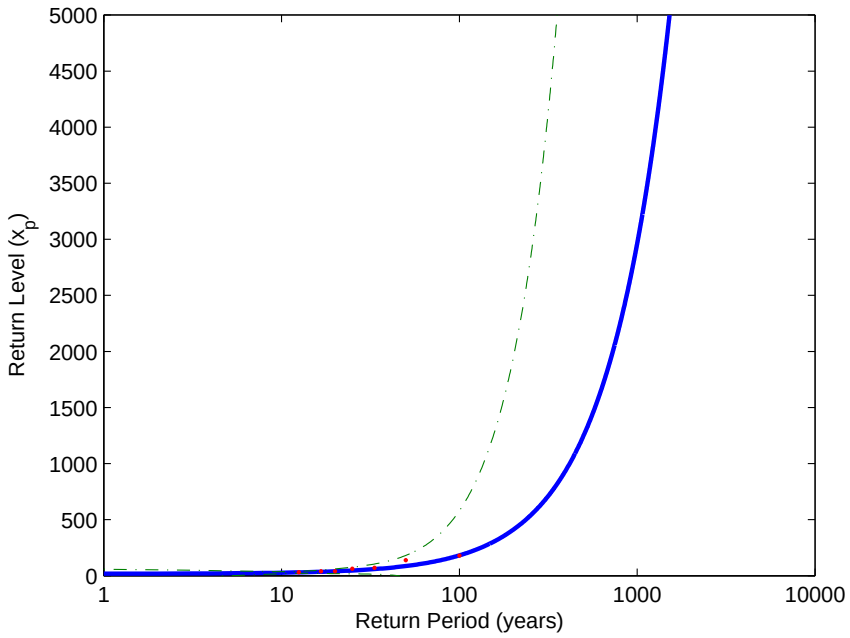
$$\mathbb{P}(X > u_0) = \frac{n}{n_y},$$

missä n on havaintojen lukumäärä, ja n_y on havaintovuosien lukumäärä. Tapahtuman $X > u$ todennäköisyys vuositasolla on nyt suoraviivaisesti

$$\mathbb{P}(X > u) = \mathbb{P}(X > u | X > u_0)\mathbb{P}(X > u_0) = \frac{k}{n_y},$$

eli otoskeskiarvo (ylityksiä per vuosi).

Tason $u = 30$ vuotuisen ylitystodennäköisyyden estimaatti yhdistetyssä aineistossa on $\hat{\lambda}_y = 13/100 = 0.13$, kun se suomalaisia koskevan aineiston perusteella on $\hat{\lambda} = 8/100 = 0.08$. Kuvaan 3.37 on piirretty toistumistasokuvaaja edellisen osion GP-ylitejakaumaan ja suomalaisia koskevan datan ylitystodennäköisyyksiin perustuen, sekä lisäksi Suomen datan havainnnot. Toistumistasokuvaajan sopivuus havaintoihin osoittautuu nyt erittäin hyväksi.



Kuva 3.37: Onnettomuuskuolemien toistumistasokuvaaja yhdistettyyn aineistoon perustuvassa GPD-mallissa ($u = 30$) Suomen dataa vastaavalla sattumistodennäköisyydellä.

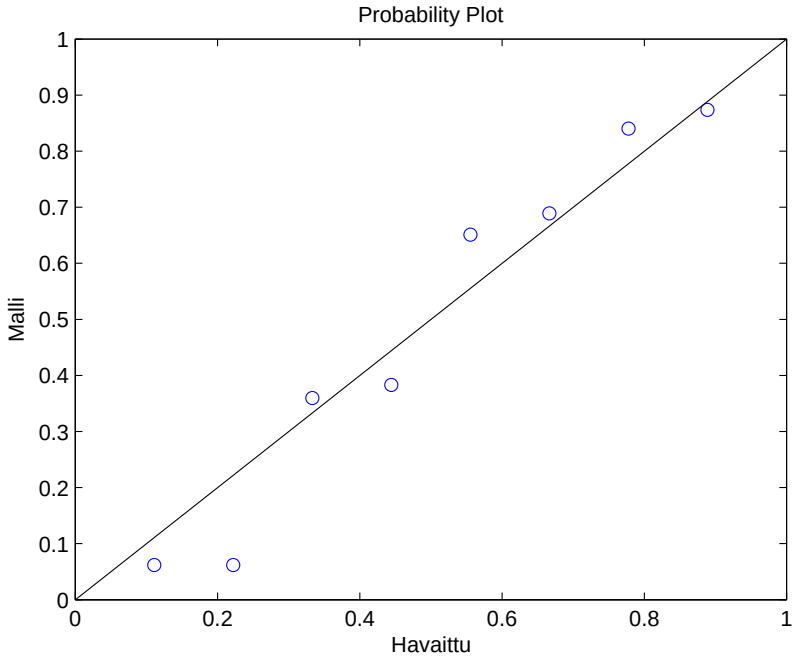
Taulukossa 3.9 alla on esitetty 10-, 100-, 1 000- ja 10 000-vuoden toistumistasoestimaatit asymptoottisine luottamusväleineen yhdistettyyn aineistoon perustuvassa GPD-mallissa Suomen dataa vastaavalla sattumistodennäköisyydellä.

Taulukko 3.9: Delta-menetelmään ja profiliuskottavuuteen perustuvat 95 %:n luottamusvälit onnettomuuskuolemien toistumistasoille yhdistettyyn aineistoon perustuvassa skaalatussa GPD-mallissa.

$u = 30$	Delta-menetelmä		Profiiliuskottavuus	
Toistumisperiodi	SUE	95% CI	SUE	95% CI
10	27	$[18, 37]$	27	-
100	180	$[-220, 580]$	180	$[75, 3500]$
1 000	3000	$[-20000, 26000]$	3000	$[280, 8.9 \cdot 10^6]$
10 000	53000	$[-660 \cdot 10^3, 770 \cdot 10^3]$	53000	$[820, 25 \cdot 10^9]$

Jos verrataan suomalaisia koskevan onnettomuuskuolemadatan empiirisiä todennäköisyyksiä GP-mallin antamiin, ja havaittuja kvantiileja mallikvantiilei-

hin, saadaan kuvat 3.38 ja 3.39. Erityisesti kvantiilikuvaajan perusteella aiemmin havaittu ongelma ylitejakauman sopimattomuudesta äärimmäisimpiin havaintoihin (ks. esimerkiksi kuva 3.17) on pitkälti poistunut.



Kuva 3.38: Todennäköisyyskuvaaja: suomalaisia koskeva onnettomuuskuolemadata vs. yhdistetty skaalattu malli; $u = 30$.

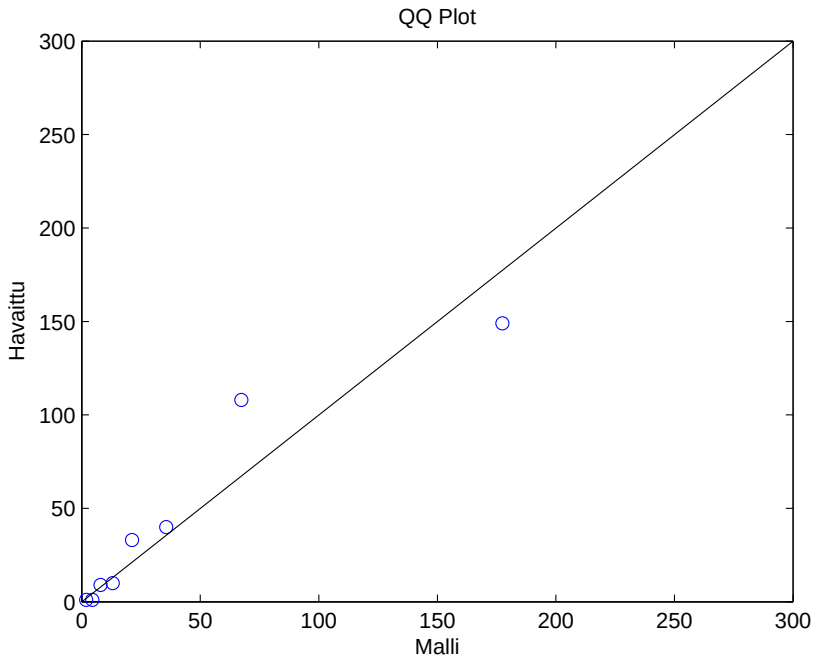
3.2.3 Pisteprosessit

Edellisten osioiden testien perusteella tason $u = 30$ (ja jo tason $u = 3$) ylityksiä tarkasteltavassa onnettomuuskuolemadatassa voidaan pitää homogeenisena Poisson-prosessina. Tämä tarkoittaa, että ylitteiden sattumisaikoja ja niiden suuruuksia voidaan tarkastella erillisinä komponentteina (ns. Poisson-GP–malli). Sovitetaan seuraavassa vertailun vuoksi pisteprosessimalli kerralla koko ylitedataan $\{(t_i, \tilde{x}_i) : i = 1, \ldots, k\}$, missä $N_u = k$ on havaittu tason u ylitysten lukumäärä. Pisteprosessilähestymistapa mahdollistaa myös epästationaaristen mallien sovittamisen dataan luontevasti, kuten aiemmin on nähty. Tutkitaan, pystytäänkö epähomogeenisen kaksiulotteisen Poisson-prosessin käytöllä parantamaan mallin sopivuutta dataan perus-POT-malliin verrattuna.

Aloitetaan sovittamalla ylitteisiin perusmuotoinen POT-malli eli ajan suhteen homogeeninen Poisson-pisteprosessi intensiteetillä

$$\lambda(t, x) \equiv \lambda(x) = \frac{1}{\sigma}\left(1 + \xi\frac{x - \mu}{\sigma}\right)^{-1/\xi - 1}$$

pisteessä $(t, x) \in E = (0, n] \times (u, \infty)$; ks. osiot 1.6 ja 2.6. Parametriestimaateiksi tason $u = 30$ ylitteitä tarkasteltaessa saadaan ainoastaan suomalaiset



Kuva 3.39: Kvantiilikuvaaja: suomalaisia koskeva onnettomuuskuolemadata vs. yhdistetty skaalattu malli; $u = 30$.

sisältävälle aineistolle

$$\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\mu}, \hat{\sigma}) = (0.731, 7.13, 3.13),$$

ja yhdistetylle aineistolle

$$\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\mu}, \hat{\sigma}) = (1.26, 18.9, 1.16),$$

Muotoparametrien ξ estimaatit ovat samat kuin edellä saatiin suoralla GP-jakauman sovittamisella, ja mallin implikoimiksi GP-jakauman skaalaparametreiksi $\hat{\beta}$ saadaan yhteyttä $\beta = \sigma + \xi(u - \mu)$ käyttämällä 19.8 ja 15.1, kuten edellä. POT-mallin maksimoiduksi log-uskottavuudeksi saadaan suomalaisten datalle -66.0 ja yhdistetylle aineistolle -104.1. Merkitään mallia $\mathcal{M}_0$.

3.2.3.1 Epähomogeeninen Poisson-pisteprosessi

Tarkastellaan ajan suhteen epähomogeenista Poisson-prosessimallia intensiteetillä

$$\lambda(t, x) = \frac{1}{\boldsymbol{\sigma}(t)} \left(1 + \boldsymbol{\xi}(t) \frac{x - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)}\right)^{-1/\boldsymbol{\xi}(t)-1},$$

missä mallin parametrisoinniksi otetaan seuraavat:

- malli $\mathcal{M}_1$: $\boldsymbol{\theta} = (\xi, \boldsymbol{\mu}(t), \sigma)$, missä $\boldsymbol{\mu}(t) = \kappa_0 + \kappa_1 t$,
- malli $\mathcal{M}_2$: $\boldsymbol{\theta} = (\xi, \mu, \boldsymbol{\sigma}(t))$, missä $\boldsymbol{\sigma}(t) = e^{\kappa_0 + \kappa_1 t}$.

Vertailukohtana käytetään edellisen osion ajan suhteen homogeenista POT-mallia, jota merkittiin $\mathcal{M}_0$. Tarkastellaan jatkossa tilan säästämiseksi vain tasoa $u = 30$ ja sen ylittävien havaintojen muodostamaa dataa.

Malli $\mathcal{M}_1$. Sovitetaan malli yhdistettyyn aineistoon käyttäen kynnystasoa $u = 30$. Suurimman uskottavuuden menetelmällä mallin parametriestimaateiksi saadaan

$$\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\kappa}_0, \hat{\kappa}_1, \hat{\sigma}) = (1.10, 20.7, -0.113, 1.86).$$

Maksimoidun log-uskottavuuden arvo on -103.8. Verrataan mallia $\mathcal{M}_1$ perusmalliin $\mathcal{M}_0$: uskottavuusosamäärätestin testisuureen arvoksi saadaan

$$D = 2\left(l_1(\mathcal{M}_1) - l_0(\mathcal{M}_0)\right) = 2(-103.8 - (-104.1)) = 0.6,$$

mikä on pieni χ^2_1-jakauman skaalalla. Esimerkiksi luottamustasoa 95 % vastaava kriittinen taso on $c_{0.05} = 3.84 > D$. Siis malli $\mathcal{M}_1$ ei paranna mallia $\mathcal{M}_0$ tilastollisesti merkitsevästi.

Pelkkään suomalaisia koskevaan aineistoon sovitettuna mallin $\mathcal{M}_1$ maksimaaliseksi log-uskottavuudeksi saadaan -65.4, jolloin testisuureen arvoksi tulee $D = 1.2 < c_{0.05}$. Siis johtopäätös on sama, trendi lokaatioparametrissa ei paranna mallia.

Malli $\mathcal{M}_2$. Maksimoimalla log-uskottavuus yhdistettyyn aineistoon sovitetun mallin parametrien SU-estimaateiksi tulee

$$\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\kappa}_0, \hat{\kappa}_1, \hat{\sigma}) = (1.28, 19.4, 0.500, -0.0091),$$

ja maksimaalinen log-uskottavuuden arvo on -103.9. Perusmalliin $\mathcal{M}_0$ verrattaessa testisuureen arvoksi saadaan näin

$$D = 2\left(l_2(\mathcal{M}_2) - l_0(\mathcal{M}_0)\right) = 2(-103.9 - (-104.1)) = 0.4.$$

Malli $\mathcal{M}_2$ ei siis paranna perusmallia. Tulos on sama vertailukohtana käytetyn vain suomalaisia koskevat onnettomuudet sisältävän aineiston kohdalla: dataan sovitetun mallin log-uskottavuus on maksimissa -65.3, ja testisuureen arvo on $D = 1.4 < c_{0.05}$.

Vertailun vuoksi edellä tarkastellut mallit sovitettiin myös tason $u = 3$ ylitteisiin, eli koko havaintoaineistoon. Trendi skaalaparametrissa σ (malli $\mathcal{M}_2$) ei näissäkään tapauksissa parantanut perusmallia tilastollisesti merkitsevästi 95 %:n luottamustasolla. Sen sijaan trendin lokaatioparametrissa μ sisältävää mallia $\mathcal{M}_1$ perusmalliin verrattaessa uskottavusosamäärätestisuureiden D arvoiksi saadaan suomalaisten aineistolle 6.8 ja yhdistetylle aineistolle 28.6. Esimerkiksi merkitsevyystasolla $\alpha = 0.01$ kriittinen arvo on $c_{0.01} = 6.63$, eli testi osoittaa vahvaa evidenssiä trendin sisältävän mallin $\mathcal{M}_1$ puolesta, ja perusmalli $\mathcal{M}_0$ tässä tapauksessa hylätään. Täytyy kuitenkin muistaa, että tason $u = 3$ ylitteisiin sovitetut mallit osoittautuivat riittämättömiksi suuronnettomuuksien kuvaamiseen. Parametrin κ_1 eli trendisuoran kulmakertoimen estimaatti on molemmissa tapauksissa positiivinen, suomalaisten aineistolle 0.080 ja yhdistetylle aineistolle 0.10. Tämä selittynee sillä, että tarkastelujakson alussa (1920-luvun tienoilla) on datassa havaittavissa pitkä ”tyhjä” jakso ilman onnettomuuksia, kun taas tarkastelujakson loppuun sijoittuu suurin tai kaksi suurinta havaittua onnettomuutta.

3.3 Onnettomuuskuolemariskin mittaamisesta

Edellä käytettiin toistumistasoja onnettomuuskuolemien jakauman hännän kuvailemiseen. Toistumistaso kuvaa suoraan ilmiöön liittyvää riskiä: esimerkiksi 100-vuoden tapahtumaa (toistumisperiodia) vastaava toistumistaso on sellainen ilmiön taso, joka odotusarvoisesti ylitetään kerran sadassa vuodessa (kun taustalla oleva prosessi on stationaarinen). Toisin ilmaistuna, tällaisen ilmiön tapahtumistodennäköisyys seuraavan vuoden aikana on 0.01. Kuten vedenkorkeuden mallintamisen yhteydessä ohimennen mainittiin, toistumistaso vastaa itse asiassa riskin mittaamiseen yleisesti käytettyä Valua-at-Risk –riskimittaa (lyh. VaR), sillä molemmat ovat yksinkertaisesti todennäköisyysjakauman kvantiileja.

Toistumistason määritelmän mukaan (määritelmä 2.2) t-vuoden toistumistaso on

$$x_t = q_{1-1/t}(F) = F^{\leftarrow}\left(1 - \frac{1}{t}\right),$$

missä $t > 0$, q on alla olevan jakauman F kvantiilifunktio ja $F^{\leftarrow}$ on jakauman F yleistetty käänteisfunktio (jatkuvalle jakaumalle $F^{\leftarrow} = F^{-1}$). Tämä on yhtä kuin Value-at-Risk luottamustasolla $\alpha = 1 - 1/t$: $\mathrm{VaR}_\alpha = q_\alpha(F) = x_t$. Esimerkiksi 100-vuoden toistumistaso vastaa (1-vuoden) VaR:ia tasolla $\alpha = 0.99$.

Toisin sanoen edellä toistumistasojen yhteydessä laskettiin implisiittisesti 1-vuoden VaR_α, $\alpha \in \{0.9, 0.99, 0.999, 0.9999\}$. Tälle voidaan johtaa GP-mallissa myös eksplisiittinen esitys osion 2.3.3 mukaisesti kääntämällä jakauma $G_{\xi,\beta}$:

$$\mathrm{VaR}_\alpha = u + \frac{\beta}{\xi}\left(\left(\frac{1-\alpha}{\bar{F}(u)}\right)^{-\xi} - 1\right),$$

kun $\alpha \geq F(u)$; ks. myös osio 4.2. Estimaatti VaR:ille saadaan, kun yo. lausekkeessa korvataan kaikki parametrit estimaateillaan, ja todennäköisyytenä $\bar{F}(u) = \mathbb{P}(X > u)$ käytetään estimaattia k/n, eli tason u ylityksien osuutta kaikista havainnoista, alaosion 3.2.2.1 mukaisesti.

Taulukossa 3.10 on esitetty onnettomuuskuolemien määrän 1-vuoden Value-at-Risk perustuen alaosion 3.2.2.1 malliin (yhdistettyyn aineistoon suomalaisten sattumisfrekvenssillä) kynnyksellä $u = 30$.

Taulukko 3.10: Onnettomuuskuolemien 1-vuoden Value-at-Risk eri tasoilla.

Luottamustaso, α	**VaR_α**
0.90	27
0.95	40
0.99	180
0.995	410
0.999	3000
0.9995	7100
0.9999	53000

Esimerkiksi Solvenssi II:ssa pääomavaatimukset lasketaan yleisesti perustuen 1-vuoden 99.5 %:n Value-at-Riskiin $\mathrm{VaR}_{0.995}$.[9] Taulukon 3.10 mukaisesti onnettomuuskuolemien määrän $\mathrm{VaR}_{0.995}$:n piste-estimaatti tarkastellussa mallissa on 410 kuollutta (ylemmän 95 % luottamusvälin riskimitan estimaatille ollessa 1900). Tällaisen tapahtuman rahallisen vaikutuksen arvioimiseksi täytyy tietysti yhdistää kuolemien määrän ja niiden vaikutusten arviointi (käytännössä yleensä simuloimalla), jotta saadaan vahinkojen kokonaisjakauma; vaikutuksen arvioinnista yleisesti ks. luvun alku.

Toisen yleisesti käytetyn riskimitan, ns. odotetun vajeen eli Expected Shortfallin (ks. osio 4.2) estimointi ei tässä tapauksessa onnistu, koska GP-jakauman muotoparametrin estimaattille pätee $\hat{\xi} > 1$, eli jakauman odotusarvo ei ole olemassa.

3.3.1 Vertailu Solvenssi II:een

Vertaillaan malliin perustuvaa 1-vuoden Value-at-Riskiä 99.5 %:n luottamustasolla – eli 410 hengen kuolemaa vastaava tapahtumaa – Solvenssi II:sen teknisissä spesifikaatioissa käytettyihin malleihin katastrofiriskistä aiheutuvan pääomavaateen laskennalle.

Vertailu tehdään viidennessä Solvenssi II:sen vaikuttavuusarvioinnissa eli ns. QIS 5:ssa (Quantitative Impact Study 5, julkaistu 7.5.2010)[10], ja kirjoitushetkellä viimeisimmässä, 28.01.2013 julkaistussa ns. LTGA (Long-Term Guarantees Assessment) –vaikuttavuusarvioinnissa[11] käytettyihin katastrofiriskispesifikaatioihin. LTGA-spesifikaation tekniset yksityiskohdat perustuvat EIOPA:n 21.12.2012 julkaisemiin uudistettuihin teknisiin spesifikaatioihin ("Revised Technical Specifications for the Solvency II valuation and Solvency Capital Requirements calculations"). Solvenssi II:sen lopulliset yksityiskohdat ovat kirjoitushetkellä (kevät 2013) vielä auki, ja niistä käydään poliittisia neuvotteluja jäsenmaiden kesken.

3.3.1.1 QIS5-spesifikaatio

QIS5:sen sairausvakuutus-moduulin katastrofi-alamoduulissa (health catastrophe risk sub-module) katastrofiriskiä mitataan Solvenssi II:sen standardikaavaa käytettäessä kolmella standardisoidulla skenaariolla:

- Areenariski (arena disaster)
- Keskittymäriski (concentration scenario)
- Pandemiariski (pandemic scenario)

[9] Vaikka katastrofiriskiä ei suoraan arvioidakaan tällä tavalla SII:n standardikaavaa käytettäessä.

[10] QIS5-määritykset löytyvät EIOPA:n (European Insurance and Occupational Pensions Authority, Euroopan vakuutus- ja lisäeläkeviranomainen) verkkosivuilta osoitteesta `https://eiopa.europa.eu/consultations/qis/insurance/quantitative-impact-study-5/index.html`.

[11] Julkaistu osoitteessa `https://eiopa.europa.eu/consultations/qis/insurance/long-term-guarantees-assessment/index.html`.

Ks. QIS5-spesifikaatio, kohta SCR.8.5. Näistä keskittymäriskiä ja erityisesti areenariskiä vastaavia skenaarioita voidaan verrata tämän luvun katastrofimalliin – pandemiariski ja taudit sen sijaan eivät sisälly malliin.

Konsentraatioriski. Konsentraatioskenaarion pyrkimyksenä on kuvata maantieteellisesti keskittynyttä vakuutettujen joukkoa kohtaavan katastrofin vaikutusta vakuutusyhtiöön; spesifikaatiossa mainitaan esimerkkinä finanssikeskuksen toimistokortteleissa sattuva suuronnettomuus. Konsentraatioriskistä aiheutuvan pääomavaateen perustana oleva tappion määrä lasketaan seuraavalla tavalla (SCR.8.128):

$$L_{co} = \sqrt{\sum\nolimits_c (L_{co,c})^2},$$

missä tietyssä maassa c sijaitsevista keskittymistä aiheutuva riski on

$$L_{co,c} = C_c \sum\nolimits_p x_p E_{c,p}. \tag{3.1}$$

Kaavassa (3.1) C_c on vakuutusyhtiön riskikonsentraation suuruus maassa c eli spesifikaation mukaan suurin tiedossa oleva yksittäisessä rakennuksessa työskentelevien vakuutettujen lukumäärä, mukaan lukien kaikki tiedossa olevat vakuutetut jotka työskentelevät 300 metrin säteellä rakennuksesta. p puolestaan tarkoittaa vakuutustuotetta (turvan tyyppiä), $E_{c,p}$ exposurea eli tässä keskimääräistä riskisummaa per henkilö turvatyypissä p, ja x_p sitä osuutta onnettomuudelle altistuneista vakuutetuista, joka kuolee tai saa turvan p kattaman vamman. Tarkasteltavat tuote- tai turvatyypit sekä näitä vastaavat vaikuttavuusasteet x_p on esitetty taulukossa 3.11 alla.

Taulukko 3.11: Sairausvakuutuksen katastrofiriskimoduulissa käytetyt tuotetyypit ja vammajakauma.

Turvatyyppi	$\boldsymbol{x_p}$ (%)
Tapaturmainen kuolema	10.0
Pysyvä täysi työkyvyttömyys	1.5
Pitkäaikainen työkyvyttömyys	5.0
Lyhytaikainen työkyvyttömyys	13.5
Sairauskulu	30.0
Yhteensä	60.0

Tarkastellaan tämän luvun sovelluksen mukaisesti tapaturmaisten kuolemien määriä. Kaavan (3.1) implikoima pääomavaateen perustana olevien onnettomuuskuolemien määrä on

$$N_{co}^{ad} = C x_{ad} = 0.1C,$$

kun vakuutusyhtiön suurin konsentraatio tarkastelumaassa (Suomessa) on C. Jos C on vaikkapa 500, saadaan konsentraatioriskin laskutavan mukaiseksi onnettomuustapahtumaksi 50 hengen kuolema (lisäksi tulevat tietysti muut loukkaantumiset). Tällainen vaikuttaa käsitellyn mallin valossa *suuruudeltaan* täysin uskottavalta. Eri asia tietysti on, kuinka todennäköistä onnettomuuden sattuminen juuri tietyn vakuutusyhtiön vakuutetuille on.

Areenariski. Areenaskenaarion pyrkimyksenä on kuvata vakuutusyhtiölle aiheutuvaa riskiä siitä, että suuri määrä ihmisiä on keskittynyt samaan aikaan samaan paikkaan, ja kyseisessä paikassa tapahtuu katastrofaalinen onnettomuus. Nimensä mukaisesti skenaarion tausta-ajatuksena on, että onnettomuus sattuu ihmisiä täynnä olevalla areenalla (esim. pommi räjähtää tai lentokone putoaa stadionille kesken suuren urheilutapahtuman tai konsertin).

Varsinaisia katastrofikuolemia ja aiheutuvaa pääomavaatimusta ajatellen areenariski lienee QIS5:n katastrofiskenaarioista merkittävin. Areenariskistä aiheutuva pääomavaade perustuu kokonaistappioon (SCR.8.114)[12]

$$L_{ar} = \sqrt{\sum\nolimits_c (L_{ar,c})^2},$$

missä maakohtainen tappion määrä on

$$L_{ar,c} = 0.5 S_c \sum\nolimits_p R_{c,p} x_p E_{c,p}. \tag{3.2}$$

Laskukaavassa p, x_p ja $E_{c,p}$ ovat kuten edellä, S_c on maakohtainen maan suurimman areenan kapasiteetti, ja $R_{c,p} = IP_{c,p}/N_c^{pop}$, missä $IP_{c,p}$ on yhtiön vakuuttamien henkilöiden lukumäärä maassa c ja tuotetyypissä p, ja N_c^{pop} on maan väestön koko (määritelty spesifikaatioon liittyvässä apuexcelissä). Turvatyypit p ja vahinkotyyppien osuudet x_p ovat samat kuin konsentraatioriskin kohdalla taulukossa 3.11. Parametreille S_c käytettävät arvot eli areenoiden kapasiteetit löytyvät QIS5-spesifikaation liitteestä L.1. Suomen tapauksessa areenana käytetään Helsingin Olympiastadionia kapasiteetilla $S = 50000$ henkeä.

Parametrien R_p ja E_p tarkoituksena on taas muuntaa tarkasteltava onnettomuus yhtiökohtaiseksi rahalliseksi tappioksi. Kaavan (3.2) mukaisesti skenaariossa oletetaan, että areena on täynnä ihmisiä (lukumäärä S) katastrofin tapahtumahetkellä ja että onnettomuus vaikuttaa 50 %:iin paikallaolijoista. Tästä epäonnisemmasta puolikkaasta 60 % kuolee tai saa erityyppisiä vammoja taulukon 3.11 mukaisin osuuksin x_p. Laskukaavan implikoima onnettomuuden koko tapaturmaisia kuolemia tarkastellessa on

$$N_{ar}^{ad} = 0.5 S x_{ad} = 0.05 S,$$

mikä Suomen kohdalla tarkoittaa siis 2500 kuolemaa (tämän lisäksi tulevat erilaiset loukkaantumiset, kuten edellisessä kohdassa). Kuolemien määrä on erittäin suuri ääriarvoteoriaan perustuvan onnettomuuskuolemamallin antamaan lukuun verrattuna, kun käytetään Solvenssi II:ssä määriteltyä riskitasoa (1-vuoden $\mathrm{VaR}_{0.995} = 410$ henkeä); 2500 henkeä vastaa itse asiassa mallin mukaan 1-vuoden Value-at-Riskiä 99.89 %:n luottamustasolla. Toisin ilmaistuna, 2500 hengen onnettomuus vastaa mallissa n. 880-vuoden tapahtumaa Solvenssi II:n vuosittaisen 99.5 %:n luottamustason implikoiman 200-vuoden tapahtuman sijaan.

Katastrofiriskiä kuvaava areenaskenaario on lähtökohdiltaan sinänsä uskottava. Ei myöskään ole epäilystäkään, etteikö 2500 ihmisen hengen vaativan katastrofin sattuminen olisi mahdollista. Avainkysymys kuitenkin on, kuinka *todennäköistä*

[12] Alkuperäistä QIS5:sen kaavaa korjattiin spesifikaation julkaisun jälkeen, ks. "Errata to the QIS5 Technical Specifications" (27.9.2010) EIOPA:n verkkosivuilla.

tällainen on. Kaikkien mahdollisten tarkastellunlaiseen onnettomuuteen johtavien tapahtumaketjujen määrittäminen, ja niiden todennäköisyyksien arviointi – kuten esimerkiksi ns. todennäköisyyspohjaisessa riskianalyysissä (PRA) ydinvoimaloiden luotettavuustarkasteluissa – on selkeästi mahdotonta. Jäljelle jää olennaisesti tilastollinen lähestymistapa riskin arviointiin, tämän luvun mukaisesti. Tältä osin voidaan todeta, että relevanttiin dataan viimeiseltä sadalta vuodelta perustuvan, ääriarvoteorian tukeman mallin mukaan areenaskenaarion suuruisen onnettomuuden tapahtuminen on selvästi Solvenssi II:ssa tavoiteltua 99.5 % tasoa epätodennäköisempää. QIS5:sen mukaisen areenaskenaarion riski vaikuttaa siis liian suurelta, ainakin Suomen osalta.

QIS5:sen skenaariopohjaista lähestymistapaa nykymuodossaan voidaan myös kritisoida sen suhteen, että konsentraatioriski ja areenariski vaikuttavat olevan enemmän tai vähemmän päällekkäisiä, eikä ole ollenkaan selvää mikä näiden yhteisvaikutuksen suhde tavoiteriskitasoon on.

3.3.1.2 LTGA-spesifikaatio

LTGA-spesifikaatiossa health-moduulin katastrofiriski-alamoduulin rakenne on sama kuin aiemmassa QIS 5:ssa, eli katastrofiriskistä aiheutuva pääomavaatimus lasketaan standardikaavaa käytettäessä kolmella standardisoidulla skenaariolla:

- Joukko-onnettomuus (mass accident)
- Keskittymäriski (concentration scenario)
- Pandemiariski (pandemic scenario)

QIS 5:n areenariskiskenaarion nimi on siis muutettu joukko-onnettomuudeksi, ja sen laskentatapaa on hieman muutettu. Konsentraatioriskiskenaario on pysynyt käytännössä samana kuin aiemmin.

Katastrofiriski-alamoduulin määrittelyihin on muutenkin tehty pieniä korjauksia ja tarkennuksia QIS 5:seen verrattuna. Esimerkiksi skenaarioiden sovellusalaa on tarkennettu (SCR.8.96):

(a) Joukko-onnettomuusskenaariota sovelletaan sairaus- ja tapaturmavakuutuksiin (health insurance) ja -jälleenvakuutuksiin, poislukien lakisääteinen tapaturmavakuutus (workers' compensation insurance) ja siihen liittyvä jälleenvakutuus.

(b) Konsentraatioskenaariota sovelletaan lakisääteiseen tapaturmavakuutukseen (workers' compensation) ja ryhmävakuutusmuotoisiin ansio- tai toimeentuloturvavakuutuksiin (group income protection), sekä näihin molempiin liittyvään jälleenvakuutukseen.

(c) Pandemiaskenaariota sovelletaan sairaus- ja tapaturmavakuutuksiin (health insurance) ja -jälleenvakuutuksiin, poislukien lakisääteinen tapaturmavakuutus (workers' compensation) ja siihen liittyvä jälleenvakutuus.

Yllä oleva jako lievittää huomattavasti QIS 5:n ongelmana ollutta massa- ja konsentraatioriskiskenaarioiden päällekkäisyyttä, vaikka skenaarioiden yhdistelmän suhde 99.5 %:n Value-at-Riskiin ei vieläkään vaikuta selvältä.

Tarkastellaan seuraavaksi LTGA-spesifikaation joukko-onnettomuusriskiä; kuten mainittua, konsentraatioriski on säilynyt olennaisesti samanlaisena kuin QIS 5:ssa edellä.

Joukko-onnettomuus. Kuten areenaskenaarioissa, joukko-onnettomuusskenaarion tavoitteena on kuvata riskiä, että paljon ihmisiä sisältävässä paikassa sattuu katastrofaalinen onnettomuus. Pääomavaatimus alimoduulille on

$$SCR_{ma} = \sqrt{\sum_c SCR_{ma,c}^2},$$

missä $SCR_{ma,c}$ on maassa c sijaitsevasta riskistä aiheutuva pääomavaatimus. Pääomavaade maan c massaonnettomuusriskille määritetään sinä yhtiön omien perusvarojen (basic own funds) vähenemänä, joka yhtiölle aiheutuu välittömästä $L_{ma,c}$:n suuruisesta tappiosta[13], missä

$$L_{ma,c} = r_c \sum_p x_p E_{c,p}.$$

Kaavassa r_c on se osuus maan c väestöstä, johon onnettomuus vaikuttaa, ja x_p se osuus onnettomuuden vaikutuspiiriin osuvista henkilöistä, jotka altistuvat tyypin p tapahtumalle, eli tapahtumalle joka korvataan tuotetyypistä p. Osuuksien r_c arvot löytyvät spesifikaation liitteestä M. Tuotteet tai tapahtumat p sekä vahinko-osuudet x_p puolestaan ovat samat kuin QIS 5:ssa käytetyt (ks. taulukko 3.11). Suure $E_{c,p}$ yo. kaavassa on turvatyypistä p maksettavien korvausten kokonaismäärä: $E_{c,p} = \sum_i SI_{p,i}$, missä $SI_{p,i}$ on korvausmäärä (tilanteesta riippuen korvausten paras estimaatti tai sopimuksen mukainen enimmäismäärä) onnettomuuden seurauksena vakuutetulle i turvasta p, ja summa on yli kaikkien vakuutettujen i jotka asuvat maassa c ja ovat vakuutettuja turvan p kattaman tapahtuman varalta.

Joukko-onnettomuusskenaarion implikoima onnettomuuden suuruus tapaturmaisia kuolemia tarkastellessa on

$$N_{ma,c}^{ad} = r_c x_{ad} N_c^{pop},$$

missä $x_{ad} = 10\%$ kuten aiemmin, ja N_c^{pop} on maan c väkiluku. Suomelle $r = 0.35\%$ ja $N^{pop} = 5.40$ miljoonaa[14], jolloin skenaarion implikoimaksi onnettomuuskuolemien lukumääräksi tulee 1890. Tämä on selvästi vähemmän kuin QIS 5:n areenariskiskenaarion 2500 henkeä, mutta yhä huomattavasti suurempi kuin EVT-mallin antama 1-vuoden 99.5 %:n VaR 410 henkeä. Mallin mukaan 1890 hengen onnettomuus vastaa 1-vuoden VaR:ia 99.86 %:n luottamustasolla, tai toisin ilmaistuna kerran 700 vuodessa sattuvaa onnettomuutta.

LTGA-vaikuttavuusarvion joukko-onnettomuusriskiin pätee QIS 5:n areenariskin kohdalla edellisen alaosion lopussa sanottu. Johtopäätös siis on, että vaikka riskiä on alennettu QIS 5:een verrattuna, vaikuttaa joukko-onnettomuusriskin koko tämän luvun analyysin valossa yhä liian suurelta (ainakin Suomen osalta) SII:n tavoiteriskitasoon nähden.

[13] Pääomavaatimus ei siis suoraan ole tappio L, vaan L:n suuruisen tappion aiheuttama vähenemä omissa perusvaroissa mahdollisen vastuuvelan muutoksen tappiota kompensoivan vaikutuksen jälkeen – esimerkiksi harkinnanvaraisten lisäetujen muutoksen seurauksena. Solvenssi II:sen pääomavaatimukset lasketaan laajemminkin tällä periaatteella (ks. LTGA-spesifikaatio, SCR.2 "Loss-absorbing capacity of technical provisions and deferred taxes").

[14] Suomen väkiluku vuoden 2011 lopussa, lähde: Tilastokeskus (`http://tilastokeskus.fi`).

3.4 Onnettomuuskuolemien simuloinnista

Tarkastellaan onnettomuuskuolemien simulointia estimoituun malliin perustuen. Koska homogeeninen Poisson-prosessimalli osoittautui riittäväksi kuvaamaan havaintoja, perustetaan simulointi Poisson-GP–lähestymistapaan eli tarkastellaan onnettomuuksien sattumista ja onnettomuuksien suuruuksia erikseen. Tämä lähestymistapa on havainnollinen simulointia ajatellen.

Merkitään i. onnettomuuden suuruutta (kuolemien lukumäärää) X_i, ja olkoon onnettomuuksien lukumäärä kiinteällä tarkasteluvälillä $(0,t]$ $N(t)$. Tällöin onnettomuuskuolemien kokonaismäärä (kokonaisvahinkomäärä) vastaavalla välillä on

$$S_N = S_{N(t)} = X_1 + \ldots + X_{N(t)} = \sum_{i=1}^{N(t)} X_i.$$

Satunnaismuuttujaa S_N kutsutaan satunnaiseksi summaksi. Kun (X_i) on jono iid satunnaismuuttujia, ja N ja (X_i) ovat riippumattomia toisistaan, kutsutaan S_N:ää edelleen yhdistetyksi summaksi tai joskus yhdistetyksi muuttujaksi.

Tarkastellussa mallissa onnettomuuksien suuruudet X_i ovat iid GP-jakautuneita satunnaismuuttujia yhteisellä kertymäfunktiolla $G = G_{\xi,\beta}$, ja vahinkojen lukumäärä $N(t)$ Poisson-jakautunut parametrilla λt, $\lambda > 0$. Suoraan mallioletuksien perusteella $N(t)$ on myös riippumaton kaikista satunnaismuuttujista $\{X_i : i = 1, 2, \ldots\}$. Näin saadun satunnaismuuttujan S jakaumaa kutsutaan yhdistetyksi Poisson-jakaumaksi; merkitään $S_N \sim \mathrm{YPoi}(\lambda, G)$.

Vahinkojen lukumäärä. Kuten edellä, otetaan perusaikayksiköksi vuosi ja tarkastellaan onnettomuuskuolemia T vuoden ajanjaksolla eli välillä $(t, t+T]$ (stationaarisuuden vuoksi voidaan ottaa yhtä hyvin väli $(0,T]$). Mallioletusten mukaisesti onnettomuuksien lukumäärä yksikön pituisella välillä $(t, t+1]$ on Poisson-jakautunut parametrilla $\lambda \geq 0$, ja siis T:n pituisella aikavälillä vastaavasti parametrilla λT. Merkitään onnettomuuksien lukumäärää $N((0,t]) := N(t) \sim \mathrm{Poi}(\lambda t)$. Siis

$$\mathbb{P}(N(t) = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \quad n = 0, 1, 2, \ldots,$$

kiinteällä t, ja sovitaan, että $\mathbb{P}(N = 0) = 1$, jos $\lambda = 0$ (tällainen degeneroitunut tapaus ei tietenkään ole kiinnostava käytännössä, ja sovelluksien kohdalla voidaan olettaa että $\lambda > 0$). Merkitään lyhyesti $p_n = \mathbb{P}(N(t) = n)$.

Yhdistetty jakauma. Yhdistetyn Poisson-jakauman kertymäfunktiolle ei ole yksinkertaista analyyttistä esitystä. Olkoon yhdistetyn muuttujan S_N kertymäfunktio F. Tämä voidaan periaatteessa kirjoittaa

$$F(x) = \mathbb{P}(S_N \leq x) = \sum_{n=0}^{\infty} \mathbb{P}(S_N \leq x | N = n)\mathbb{P}(N = n) = \sum_{n=0}^{\infty} p_n G^{n*}(x),$$

missä G^{n*} on G:n n. konvoluutio, eli

$$G^{n*}(x) = \mathbb{P}(S_n \leq x) = \int_{-\infty}^{\infty} G^{(n-1)*}(x-y)\, \mathrm{d}G(y), \quad n = 1, 2, \ldots,$$

ja $G^{0*}(x) = \mathbb{1}_{\{x \geq 0\}}$. Teoriassa kertymäfunktio F on mahdollista määrätä konvoluutioiden kautta, kun tapahtumien pistetodennäköisyydet ja suuruusjakauman kertymäfunktio tunnetaan. Käytännössä tämä on työlästä, mutta jakaumasta voidaan tuottaa realisaatioita helposti simuloimalla.

Huomautus 3.1 *Olkoon S_N yhdistetty summa edellä esitetyin oletuksin ((X_i) jono iid satunnaismuuttujia sekä N ja (X_i) riippumattomia toisistaan). Tällöin S_N:n Laplace-Stieltjes–muunnokselle yleisessä tapauksessa pätee*

$$\hat{F}(s) = \int_0^\infty e^{-sx} \, \mathrm{d}F(x) = \sum_{n=0}^{\infty} p_n \hat{G}^k(s) = M_N\left(\hat{G}(s)\right), s \geq 0,$$

missä M_N on on satunnaismuuttujan N momentit generoiva funktio, $M_N(s) = \mathbb{E}(e^{sN})$. Kun N on Poisson-jakautunut, $N \sim \mathrm{Poi}(\lambda)$, momentit generoiva funktio on

$$M_N(s) = \mathbb{E}(e^{sN}) = \sum_{n=0}^{\infty} p_n e^{sn} = \sum_{n=0}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} e^{sn} = e^{-\lambda(1-s)}.$$

Tarkasteltavassa Poisson-GP-mallissa siis yhdistetylle jakaumalle saadaan

$$\hat{F}(s) = M_N\left(\hat{G}(s)\right) = \exp\left(-\lambda(1 - \hat{G}(s))\right), \quad s \geq 0.$$

$\mathrm{YPoi}(\lambda, G)$-jakautuneiden satunnaislukujen simulointi perustuu seuraavaksi esitettävään tulokseen. Oletetaan, että käytössä on haluttu määrä riippumattomia välille $(0, 1)$ tasajakautuneita satunnaismuuttujia, ja merkitään geneeristä tällaista muuttujaa $U \sim T(0, 1)$. Kertymäfunktiolle T siis pätee

$$T(x) = \begin{cases} 0, & x < 0, \\ x, & x \in [0, 1], \\ 1, & x > 1. \end{cases}$$

Lemma 3.2 (Käänteismuunnosmenetelmä)
Olkoon F kertymäfunktio ja $U \sim T(0, 1)$. Asetetaan

$$U_F = \inf\{x \in \mathbb{R} : F(x) \geq U\} = F^{\leftarrow}(U).$$

Tällöin U_F noudattaa jakaumaa F, eli

$$\mathbb{P}(U_F \leq x) = F(x), \quad \forall\, x \in \mathbb{R}.$$

Koska funktio F on kertymäfunktiona oikealta jatkuva, pätee kaikilla $x, y \in \mathbb{R}$

$$\begin{cases} F^{\leftarrow}(y) \leq x & \iff \quad y \leq F(x) \\ x < F^{\leftarrow}(y) & \iff \quad y > F(x) \end{cases} \tag{3.3}$$

yleistetyn käänteisfunktion $F^{\leftarrow}$ määritelmän perusteella. Lemman 3.2 tulos seuraa nyt suoraviivaisesti: kun $x \in \mathbb{R}$,

$$\mathbb{P}(U_F \leq x) = \mathbb{P}(F^{\leftarrow}(U) \leq x) = \mathbb{P}(U \leq F(x)) = T(F(x)) = F(x).$$

Lauseen tulos pätee sekä diskreeteille että jatkuvilla jakaumille, ja näiden sekoituksille. Jatkuville jakaumille pätee edelleen $F^{\leftarrow}(x) = F^{-1}(x)$, $\forall\, x \in \mathbb{R}$, eli voidaan suoraan asettaa $U_F = \min\{x \in \mathbb{R} : F(x) = U\} = F^{-1}(U) \sim F$.

Kuten edellä mainittiin, yhdistetyn Poisson-jakauman kertymäfunktiolle ei ole olemassa yksinkertaista analyyttista esitystä, jolloin lemmaa 3.2 ei voida soveltaa suoraan. Yhdistettyä Poisson-jakaumaa noudattavien satunnaismuuttujien generointi onnistuu kuitenkin seuraavalla kaksivaiheisella menettelyllä. Olkoon tarkasteltavan aikavälin pituus $T > 0$ vuotta.

Algoritmi 3.3 (YPoi-jakautuneiden satunnaismuuttujien generointi)

1. Generoidaan Poisson-jakautunut satunnaisluku parametrilla λT, merkitään N_i, seuraavasti:
   a) Generoidaan tasajakautunut satunnaismuuttuja $U \sim T(0,1)$.
   b) Asetetaan $N_i = k$, jos $F_{\mathrm{Poi}}(k-1) < U \leq F_{\mathrm{Poi}}(k)$; F_{Poi} on Poisson-jakautuneen satunnaismuuttujan kertymäfunktio,
   $$F_{\mathrm{Poi}}(x) = \sum_{j=1}^{\lfloor x \rfloor} e^{-\lambda T} \frac{(\lambda T)^j}{j!},$$
   missä $\lfloor \cdot \rfloor$ tarkoittaa kokonaislukuosaa.
2. Generoidaan N_i kpl iid satunnaislukuja yleistetystä Pareto-jakaumasta $G_{\xi,\beta}$, merkitään näitä $X_1, \ldots, X_{N_i}$:
   a) Generoidaan tasajakautunut satunnaismuuttuja $U \sim T(0,1)$.
   b) Jos $\xi = 0$, asetetaan
   $$\tilde{X}_j = -\ln U,$$
   muutoin (kun $\xi \neq 0$)
   $$\tilde{X}_j = \frac{1}{\xi}(U^{-\xi} - 1).$$
   c) Asetetaan edelleen $X_j = u + \beta \tilde{X}_j$, missä u on GP-jakauman $G_{\xi,\beta}$ sovituksessa käytetty kynnystaso. X_j noudattaa nyt jakaumaa $G_{\xi,\beta}$.
   d) Toistetaan menettelyä kunnes on saatu N_i satunnaislukua, $j = 1, \ldots, N_i$.
3. Asetetaan $S_i = X_1 + \cdots + X_{N_i}$. Näin saatu S_i noudattaa haluttua yhdistettyä Poisson-jakaumaa.

Toistamalla menettely n kertaa saadaan iid otos $(S_i : i = 1, \ldots, n)$ yhdistetystä jakaumasta. Olennaista yllä on, että kaikki esiintyvät satunnaismuuttujat pidetään toisistaan riippumattomina.

Huomautus 3.4 *Edellä nähtiin, että N:n riippumattoman GP-jakautuneen satunnaismuuttujan summan jakauman määrittäminen vaatii konvoluution laskemista. N-havainnon maksimin jakauma sen sijaan voidaan kirjoittaa helposti eksplisiittisessä muodossa, kuten osiossa 1.6 nähtiin. Merkitään maksimia $M_N = \max(X_1, \ldots, X_N)$, missä siis $N \sim \mathrm{Poi}(\lambda)$ ja on riippumaton iid satunnaismuuttujajonosta (X_n) GP-jakaumalla $G_{\xi,\beta}$. Nyt pätee*

$$\mathbb{P}(M_N \leq x) = \exp\left\{ -\left(1 + \xi \frac{x-\mu}{\beta}\right)^{-1/\xi} \right\} = H_{\xi,\mu,\sigma}(x),$$

missä $\mu = \frac{\beta}{\xi}(\lambda^{\xi} - 1)$ *ja* $\sigma = \beta\lambda^{\xi}$*; ks. yhtälö 1.27. Tarkastellussa mallissa ylitteiden maksimi on siis GEV-jakautunut.*

GEV-jakautuneita satunnaismuuttujia voidaan generoida seuraavasti lemmaan 3.2 perustuen:

Algoritmi 3.5 (GEV-jakautuneiden satunnaismuuttujien generointi)

1. Generoidaan tasajakautunut satunnaismuuttuja $U \sim T(0, 1)$.
2. Jos $\xi = 0$, asetetaan
$$\tilde{M}_j = -\ln(\ln U),$$
muutoin (kun $\xi \neq 0$)
$$\tilde{M}_j = \frac{1}{\xi}((-\ln U)^{-\xi} - 1).$$
3. Asetetaan $M = \mu + \beta\tilde{M}_j$. M on nyt GEV-jakautunut, $M \sim H_{\xi,\mu,\sigma}$.

3.5 Johtopäätökset

Tässä luvussa tarkasteltiin tilastoaineistoa suomalaisia kohdanneista onnettomuuksista, ja toisaalta vastaavan ruotsalaisia koskevan datan yhdistämistä tähän analyysissä käytettävissä olevan havaintoaineiston laajentamiseksi. Tavoitteena oli tutkia, mitä onnettomuuskuolemien määrän todennäköisyysjakaumasta voidaan sanoa ääriarvoteoriaan perustuvia tilastollisia menetelmiä soveltamalla.

Tuloksia tulkittaessa tulee muistaa, että tarkastelu koski nimenomaan onnettomuuskuolemia, ja esimerkiksi epidemia- ja pandemiakuolemia ei tarkasteltu. Myös sodat jätettiin tarkastelun ulkopuolelle.

Kriittisesti tarkasteltuna johtopäätöksenä on oikeastaan, että *todellisista* katastrofitason onnettomuuskuolemista ei voida sanoa kovin paljoa minkäänlaisella varmuudella pelkästään käytettävissä olevien havaintojen perusteella, äärihavaintojen niukkuudesta johtuen. Tämä ei tietenkään ole mikään yllätys ilmiön (onnettomuuskuolemien) luonnetta ajatellen. Tiedetään, että merkittävästi havaittuja suurempien onnettomuuskuolemamäärien sattuminen on mahdollista. Käytännössä puhtaasti tilastollista analyysia tulee täydentää muun ilmiön luonteesta tiedetyn teoreettisen ja kokemusperäisen tiedon avulla, esimerkiksi tarkastelemalla erilaisia mahdollisia tulevaisuuden skenaarioita. Todella äärimmäisten onnettomuuksien mahdollisuus, todennäköisyys ja vaikutukset joudutaan arvioimaan joka tapauksessa erikseen (mikäli niitä ylipäätään arvioidaan), ja päättämään niiden vaikutuksilta suojautumisesta tai suojautumatta jättämisestä vakuutusyhtiön altistumisen, riskinkantokyvyn ja riskinottohalukkuuden perusteella.

Jotain ilmiöstä voidaan silti sanoa tilastollisen mallinnuksen perusteellakin, erityisesti hieman alemmilla luottamustasoilla. Perustamalla todennäköisyyksien ja suuruusluokkien arviointi järkevään ääriarvoteorian tukemaan rajamal-

liin, voidaan ilmiöön liittyvää epävarmuutta arvioida perustellulla ja rehellisellä tavalla. Tarkasteltu yleistettyyn Pareto-jakaumaan perustuva malli mahdollistaa ekstrapolaation havaintojen ulkopuolelle, tarjoten hyödyllisen työkalun päätöksenteon tueksi: vaikka päätöksiä ei täysin suoraviivaisesti perustettaisikaan malliin, antaa tilastollinen malli kuitenkin usein korvaamattoman kuvauksen reaalimaailman ilmiöistä vähäisellä subjektiivisuuden asteella. Erilaiset simulaatiotarkastelut ja stressitestit voidaan myös toteuttaa suoraviivaisesti käsitellyn mallin perusteella, ja näiden vaikutusta vakuutusyhtiöön arvioida.

Esimerkkinä esitettiin Solvenssi II:sen käyttämän riskimitan, vuositason 99.5 %:n Value-at-Riskin estimointi onnettomuuskuolemamäärille mallia käyttäen, ja tätä verrattiin QIS5- ja LTGA-spesifikaatioissa käytettyihin katastrofiriskiä mittaamaan pyrkiviin skenaarioihin. Konsentraatioriskiskenaarion suuruus havaittiin periaatteessa uskottavaksi. Sen sijaan areenaskenaarion riski QIS 5:ssa ja tämän korvaajan, joukko-onnettomuusskenaarion riski LTGA:ssa vaikutti selvästi liian suurelta Solvenssi II:n tavoittelemaa, pääomavaatimusten määrittämisessä käytettyä riskitasoa ajatellen.

Yhteenvetona voidaan jo aiemmin sanottua toistaen todeta, että tarkasteltu malli antaa teoreettisesti perustellun tavan arvioida onnettomuuskuolemien määrää ja näihin liittyviä todennäköisyyksiä — tietyssä mielessä niin hyvin kuin havaitun datan perusteella on mahdollista. Sinällään se on käyttökelpoinen työkalu, vaikka onnettomuuskuolemien jakauman häntään suuri epävarmuus liittyykin.

Luku 4

Markkinariskin mallinnus ääriarvoteoriaa käyttäen

Tarkastellaan seuraavassa ääriarvoteorian soveltamista finanssiaikasarjoihin ja erityisesti markkinariskin mallintamiseen. Markkinariskillä tarkoitetaan tappion mahdollisuutta, joka johtuu muutoksista sijoitusinstrumenttien hinnoissa tai instrumenttien hintoihin vaikuttavissa alla olevissa tekijöissä (osakkeiden ja hyödykkeiden hinnat, korot, valuuttakurssit, jne.).

Mallinnuksen näkökulmasta riskejä edustavat satunnaismuuttujat, jotka kuvaavat tulevaisuuden maailmantilat tappioita ja voittoja edustaviksi lukuarvoiksi. Markkinariskin mallintamisessa pyrkimyksenä on mallintaa tappio- tai voitto-tappio–jakauman (profit & loss distribution, P&L) häntiä – ja jakauman häntiä tarkastellessa ollaan suoraan ääriarvoteorian sovellusalueella. Ääriarvot ovat läsnä kaikilla riskienhallinnan osa-alueilla: riittävän riskienhallintajärjestelmän on välttämätöntä käyttää malleja, jotka huomioivat harvinaiset mutta seuraamuksiltaan vakavat tapahtumat, ja mahdollistavat tällaisten tapahtumien seurausten mittaamisen ja arvioinnin.

4.1 Finanssiaikasarjojen piirteistä

Merkitään finanssi-instrumentin hinta-aikasarjaa (S_t), jolloin vastaavat logaritmiset tuotot ovat

$$r_t = \ln\left(\frac{S_t}{S_{t-1}}\right) = \ln S_t - \ln S_{t-1}.$$

Taustalla oleva hintaprosessi $(S_t)_{t\geq 0}$ on periaatteessa jatkuva-aikainen, mutta sovelluksia ajatellen keskitytään tarkastelemaan tästä tasavälein havaittujen arvojen muodostamaa prosessia $(S_t)_{t=0}^n$, $n \in \mathbb{N}$, tyypillisesti päivän päätöshintoja, jolloin sarja $(r_t)_{t=1}^n$ vastaa päivittäisiä (log-)tuottoja.

Logaritmimuunnoksen (tai yleisemmin sopivien differenssien) soveltaminen hinta-aikasarjaan tekee tuloksena olevasta tuottoaikasarjasta (r_t) likimain stationaarisen. Tyypillinen finanssimarkkinoilla havaittu tuottodata osoittaa kuiten-

kin vahvaa evidenssiä iid-oletusta vastaan. Empiirisistä tutkimuksista tiedetään, että tuottodatalla on usein seuraavat ominaispiirteet (ks. esim. [33]):

- Data on paksuhäntäistä.
- Tuottoaikasarjan autokorrelaatio on hyvin matalaa.
- Tuottoaikasarjan arvojen neliöiden tai itseisarvojen autokorrelaatio on korkeaa.
- Volatiliteetti muuttuu (satunnaisesti) ajan suhteen.
- Volatiliteetti kasaantuu eli itseisarvoltaan suuret arvot sattuvat ryppäissä.

Näitä piirteitä kutsutaan joskus finanssidataa koskeviksi tyylitellyiksi faktoiksi (*stylized facts*). Finanssi-instrumenttien hintaprosesseja kuvaavien mallien tulisi pyrkiä ottamaan nämä empiiriset löydöt huomioon. Erityisesti volatiliteetin kasaantumisen (volatility clustering) mallintaminen on tärkeää, sillä useimmat muista havainnoista voidaan selittää kokonaan tai osittain tähän ominaisuuteen perustuen.

Finanssidatan ilmentämä riippuvuusrakenne tekee tavanomaisista, iid datalle formuloiduista ääriarvoteorian tilastollisista menetelmistä pitkälti soveltumattomia.[1] Esimerkiksi ylitemenetelmä tai perus-POT-malli eivät suoraan sovellu tyypilliseen dataan volatiliteetin kasaantumisen ja sen aiheuttaman ääriarvojen ryppäissä esiintymisen vuoksi. Aiemmista luvuista muistetaan, että standardin POT-mallin ja (implisiittisesti) ylitemenetelmän taustalla on oletus, että korkean tason ylitteet tapahtuvat homogeenisen Poisson-prosessin mukaisesti. Tämä ei tavallisesti päde finanssidatan kohdalla, vaan ylityksillä on taipumus esiintyä klustereissa, vastaten ajanjaksoja jolloin volatiliteetti on korkealla.

Suoraviivainen ratkaisu klusteroitumisongelmaan on edetä deklusteroimalla data eli erottelemalla havaintoklusterit toisistaan jonkin säännön mukaisesti, ja tarkastelemalla vain klusterimaksimeja (ks. osio 2.4). Tämä ei kuitenkaan ratkaise ongelmaa sinänsä, vaan ainoastaan kiertää sen. Esimerkiksi juuri markkinariskin kohdalla lyhyen aikavälin riippuvuusrakenteen (volatiliteetin kasautumisen) ja prosessin dynamiikan mallintaminen on kiinnostavaa, ellei jopa välttämätöntä. On siis tarpeen löytää ilmiötä paremmin kuvaava mallirakenne.

Finanssidatan ääriarvojen mallintamiseen markkinariskin viitekehyksessä on esitetty kaksi pääasiallista lähestymistapaa: ns. itseherätteisten (self-exciting) pisteprosessien käyttö, sekä kaksivaiheinen menetelmä, jossa dataan sovitetaan ensin sopiva volatiliteettirakenteen huomioiva aikasarjamalli (tyypillisesti GARCH-perheestä), ja ylitemenetelmää sovelletaan näin saatuihin residuaaleihin. Ensimmäinen lähestymistapa on esitetty paperissa [34] (ks. myös [5, kappaleet 7.4.3 –4]; itseherättävillä pisteprosesseilla on alkunsa seismologisissa sovelluksissa, ks.

[1] Toki esimerkiksi blokkimaksimimenetelmää voidaan soveltaa vaikkapa päivittäisistä tuotoista poimittuihin vuosi- tai kvartaalimaksimeihin (merenpinnan korkeuden tapaan), sillä näitä havaintoja voidaan yleensä pitää riittävän riippumattomina. Tämä lähestymistapa voi sopia, kun mietitään esimerkiksi mahdollisia katastrofiskenaarioita stressitestejä varten, mutta markkinariskin mallintamisen osalta lähestymistavan hyöty on yleensä rajoitettu.

[35] ja [31]. Toisen lähestymistavan esittivät alun perin McNeil ja Frey paperissa [36].

Tarkastellaan tässä luvussa [36]:n mukaista GARCH-EVT–mallia markkinariskin mallintamiseen. Mallin ideana on siis käyttää GARCH-tyyppistä aikasarjamallia tuottoaikasarjan volatiliteetin estimoimiseen, ja soveltaa ääriarvoteoriaa tuloksena saatavan residuaalien jakauman häntiin mallintamalla niitä GP-jakaumilla.

4.1.1 Stationaarinen ja ehdollinen tuottojakauma

Kun tuottoprosessin volatiliteetti on stokastinen (eli tuotot eivät ole iid), voidaan erottaa kaksi eri tuottojakaumaa, nimittäin stationaarinen ja ehdollinen. Stationaarinen jakauma on nimensä mukaisesti prosessin ehdollistamaton (unconditional) pitkän aikavälin jakauma, kun aikasarjan (X_t) oletetaan muodostavan stationaarisen prosessin. Ehdollinen tuottojakauma puolestaan on tuottojen X_t jakauma hetkellä t ehdollistettuna prosessin historialla $\mathcal{F}_t = \sigma(\{X_s | s \leq t\})$, eli hetkeen t mennessä kertyneellä informaatiolla. Ehdollinen jakauma siis heijastelee tarkasteluhetkellä vallitsevia olosuhteita.

Stationaariseen tuottojakauman perustuvaa lähestymistapaa käytetään yleensä tarkastellessa pitkiä aikavälejä, ja esimerkiksi luottoportfolioiden hallinnassa ([5]). Ehdollinen lähestymistapa vaatii riskitekijöiden (tuottojen/tappioiden) *dynamiikan* mallintamista, ja sopii markkinariskin mittaamiseen: markkinariskin hallinnan näkökulmasta kiinnostus kohdistuu yleensä epäsuotuisista markkinaliikkeistä seuraavan raportointiperiodin kuluessa mahdollisesti aiheutuviin tappioihin, markkinoilla tarkasteluhetkellä vallitseva volatiliteettiympäristö huomioiden.

4.2 Riskimitoista

Tyypillisesti portfolion, position tai yksittäisen instrumentin riskiä mitataan jollakin asianomaisen tappiojakauman häntään liittyvällä riskimitalla, joka tiivistää riskin yhdeksi riskillisyyttä kuvaavaksi numeroksi. Olkoon aikavälin $[t, t+1]$ tappiota kuvaava satunnaismuuttuja $L_{t+1} = -r_{t+1}$. Oletetaan seuraavassa, että lähestymistapa riskin tarkasteluun (stationaarinen tai ehdollinen tuottojakauma) on jo valittu, ja merkitään tappion $L := L_{t+1}$ jakaumaa $F_L(l) = \mathbb{P}(L \leq l)$.

Ensimmäisenä laajempaan käyttöön pankkimaailmassa otettu, ja yhä yleisimmin käytetty riskimitta on Value-at-Risk (VaR), joka kuvaa "tappiota, jota ei suurella todennäköisyydellä ylitetä". Täsmällisemmin, olkoon $0 < \alpha < 1$. VaR luottamustasolla α on pienin reaaliluku l, jolle todennäköisyys että tappio L ylittää arvon l on pienempi tai yhtä suuri kuin $1 - \alpha$:

$$\mathrm{VaR}_\alpha = \inf\{l \in \mathbb{R} : \mathbb{P}(L > l) \leq 1 - \alpha\} = \inf\{l \in \mathbb{R} : F_L \geq \alpha\} = F_L^{\leftarrow}(\alpha). \quad (4.1)$$

VaR_α on siis tappiojakauman α-kvantiili, $\mathrm{VaR}_\alpha = q_\alpha(F_L) = q_\alpha(L)$. Usein on tarpeen eksplisiittisesti ilmaista myös tarkastelujakso (holding period), jota ris-

kimitta koskee. Olkoon tämä h, jolloin tarkastellaan siis aikavälin $[t, t+h]$ tappioita $L_{t+h}^{(h)} = L_{t+1} + \cdots + L_{t+h}$[2], ja merkitään näiden jakaumaan perustuvaa h-periodin Value-at-Riskiä hetkellä t $\mathrm{VaR}_\alpha^t(h)$, tai yksinkertaisemmin $\mathrm{VaR}_\alpha(h)$ jos tarkasteluhetkestä ei ole epäselvyyttä.

VaR käsitteenä on helposti ymmärrettävä (vaikkakin se silti joskus ymmärretään väärin[3]), helposti kommunikoitavissa, ja yleensä helposti laskettavissa (tehtäessä normaalisuusoletus tarvitaan olennaisesti vain arvio keskihajonnasta). VaR:iin riskin mittana liittyy kuitenkin teoreettisia ja käytännöllisiä ongelmia (ks. esim. [5]); VaR ei esimerkiksi ole koherentti riskimitta Artzner et. al.:in mielessä [37].

Toinen yleisesti käytetty riskimitta on niin sanottu odotettu vaje eli Expected Shortfall (ES). Nimensä mukaisesti sen voidaan ajatella kuvaavan keskimääräistä (odotusarvoista) tappiota siinä tapauksessa, että määrätty korkea tappiotaso (VaR) ylitetään. Tarkemmin, ES luottamustasolla $\alpha \in (0,1)$ määritellään

$$\mathrm{ES}_\alpha = \frac{1}{1-\alpha} \int_\alpha^1 q_u(F_L)\, \mathrm{d}u, \tag{4.2}$$

kun $\mathbb{E}(|L|) < \infty$ eli L on integroituva satunnaismuuttuja. Jos tappiojakauma F_L on jatkuva, ES on odotettu tappio ehdolla että vastaavan luottamustason VaR ylitetään:

$$\mathrm{ES}_\alpha = \frac{1}{1-\alpha}\mathbb{E}\left(L\mathbb{1}_{\{L \geq q_\alpha(L)\}}\right) = \frac{1}{1-\alpha}\mathbb{E}\left(L; L \geq q_\alpha(L)\right) = \mathbb{E}(L | L \geq \mathrm{VaR}_\alpha).$$

Epäjatkuville jakaumille lausekkeesta tulee monimutkaisempi,

$$\mathrm{ES}_\alpha = \frac{1}{1-\alpha}\left(E\left(L; L \geq q_\alpha(L)\right) + q_\alpha(L)\left[(1-\alpha) - \mathbb{P}(L \geq q_\alpha(L))\right]\right),$$

ks. [38]. Jatkuville jakaumille $\mathbb{P}(L \geq q_\alpha(L)) = 1-\alpha$, ja jälkimmäinen termi yllä häviää.

ES on koherentti riskimitta, toisin kuin VaR, ja voidaan osoittaa että sillä on hyvät teoreettiset ominaisuudet. Käytännössä Expected Shortfallin käyttö riskimittana voi kuitenkin olla ongelmallista, sillä ES riippuu erityisen vahvasti jakauman äärimmäisestä hännästä.[4] Tämä herkkyys itsessään on tietysti toivottava ominaisuus riskimitalta, mutta voi aiheuttaa ongelmia kun riskimitan arvo täytyy estimoida vähien äärihavaintojen perusteella. ES:n estimointi voikin

[2] Periodituottojen summautuvuus pätee koska tarkastellaan logaritmisia tuottoja: $L_{t+h}^{(h)} = -\ln\left(\frac{S_{t+h}}{S_t}\right) = -\ln\left(\frac{S_{t+h}}{S_{t+h-1}} \cdots \frac{S_{t+1}}{S_t}\right) = \sum_{j=1}^h L_{t+j}$. Tämä ominaisuus onkin yksi logaritmisten tuottojen käytön suurimpia etuja.

[3] Tyypillisin Value-at-Riskiä koskeva virheellinen tulkinta koskenee käsitteen epämuodollista määritelmää muodossa (olettaen havainnollisuuden vuoksi, että $\alpha = 0.05$) "VaR on suurin tappio, joka tapahtuu 95 %:n todennäköisyydellä / kerran 20 vuodessa...". Kuten määritelmästä nähdään, VaR on itse asiassa *pienin* tappio joka tapahtuu 5 %:ssa huonoimmista tulemista.

[4] Toisena haittapuolena voidaan nähdä, että ES on hyvin määritelty vain, kun alla olevalla jakaumalla on äärellinen ensimmäinen momentti. Vakuutussovellusten yhteydessä data viittaa joskus siihen, ettei jakauman ensimmäinenkään momentti ole olemassa; vrt. onnettomuuskuolemien määrän tarkasteluun luvussa 3.

olla vaikeaa, ja tuloksena saatavan estimaattorin varianssi on usein erittäin suuri. Estimaatti saattaa heilahdella paljon yksittäisenkin häntähavainnon muuttuessa. Tällaisessa tilanteessa soveltuvien ja perusteltujen tilastollisten menetelmien (ääriarvoteorian) käyttö on korostuneen tärkeää. Tosin havaitun datan ulkopuolelle ekstrapoloitaessa täytyy aina noudattaa varovaisuutta, ja pitää mielessä, että hienoinkin tilastollinen malli perustuu pohjimmiltaa havaitun otoksen sisältämiin havaintoihin.

Esimerkkeinä riskimittojen käytöstä, pankkeja koskevassa ns. Basel II-säännöstössä vaaditaan, että pankin riskipääoman täytyy olla riittävä kattamaan pankin kaupankäyntiportfoliosta (trading book) 10 päivän ajanjaksolla syntyvät tappiot 99 %:n todennäköisyydellä (ks. [5]); käytetty riskimitta on siis 10 päivän 99 %:n VaR eli $\mathrm{VaR}_{0.99}(10)$. Sisäisessä riskikontrollissa ja riskien raportoinnissa käytetään tavanomaisesti yhden päivän holding periodia ja 95 % luottamustasoa vastaavaa VaR:ia. Euroopan unionin jäsenmaiden vakuutusyhtiöitä koskevan Solvenssi II –vakavaraisuusdirektiivin määrittelyissä riskin mittana pääomavaatimusta asetettaessa käytetään 99.5 %:n VaR:ia vuoden ajanjaksolla. Sveitsissä puolestaan on käytössä oma riskiperusteinen vakuutusyhtiöitä koskeva vakavaraisuussäännöstönsä (Swiss Solvency Test), jossa riskimittana käytetään 99 %:n Expected Shortfallia vuoden ajanjaksolla.[5]

4.3 Dynaaminen EVT-malli finanssiaikasarjoille

Tarkastellaan tuotto- tai tappiojakauman hännän mallinnusta ja häntään perustuvien riskimittojen (VaR, ES) estimointia ääriarvoteorian avulla, yleistettyä Pareto-jakaumaa käyttäen. Ylitemenetelmä perusmuodossaan tuottodataan sovellettuna koskee stationaarista tuottojakaumaa, eli on ehdollistamaton (unconditional) menetelmä. Markkinariskin mallintamiseksi lyhyellä aikavälillä on kuitenkin tarpeen tarkastella ehdollista tuottojakaumaa, missä ehdollistus on vallitsevan (stokastisen) volatiliteettirakenteen suhteen. McNeil ja Frey [36] esittivät menetelmän, jossa yhdistetään tuottoprosessin volatiliteettirakenteen estimoimiseen käytetty (AR-)GARCH-malli GP-jakaumaan, jolla mallinnetaan aikasarjamallin innovaatiojakauman hännät.

Oletetaan seuraavassa, että (logaritmiset) tuotot noudattavat stationaarista aikasarjaprosessia stokastisella volatiliteettirakenteella, ja tarkastellaan päivätuottoja. Olkoon $(X_t)_{t\in\mathbb{Z}}$ negatiivisten log-tuottojen muodostama aikasarja (yksinkertaisuuden vuoksi merkitään $X_t = L_t$). Prosessin oletetaan olevan muotoa

$$X_t = \mu_t + \sigma_t Z_t,$$

missä $\mu_t, \sigma_t \in \mathcal{F}_{t-1}$, ja innovaatiot (Z_t) ovat iid satunnaismuuttujajono[6] jollakin (tuntemattomalla) jakaumalla G.

[5] Ks. esim. FINMA:n (Swiss Financial Market Supervisory Authority) verkkosivut, `http://www.finma.ch/e/beaufsichtigte/versicherungen/schweizer-solvenztest/Pages/default.aspx`.

[6] Riippumattomien ja samoin jakautuneiden satunnaismuuttujien muodostamaa prosessia kutsutaan aikasarjakontekstissa usein myös valkoiseksi kohinaksi (white noise, WN).

Prosessin (X_t) ehdollinen odotusarvo μ_t ja volatiliteetti σ_t mallinnetaan käyttäen ARMA- ja GARCH-tyyppisiä aikasarjamalleja. Mallit sovitetaan käyttäen kvasi-SU-menetelmää (quasi-maximum likelihood, QML), koska innovaatiojakaumasta G tehdään mahdollisimman vähän oletuksia tässä vaiheessa. Tavoiteltua markkinariskisovellusta ajatellen innovaatiojakauman hännät mallinnetaan yleistetyllä Pareto-jakaumalla; koska innovaatiojakaumaa ei havaita suoraan, käytetään ensimmäisen vaiheen mallin sovituksesta saatuja residuaaleja lähtödatana.

Luodaan ensin hyvin lyhyt katsaus ARCH/GARCH-malliperheeseen, minkä jälkeen tarkastellaan GARCH-tyypin mallin sovittamista tuottodataan sekä GP-jakauman sovitusta tuloksena saadun residuaalijakauman häntiin. Tämän jälkeen yhdistetyn mallin perusteella esitetään menetelmä riskimittojen estimoimiseksi, ja menetelmää sovelletaan osakemarkkinadataan.

4.3.1 ARCH/GARCH–malliperhe

Olkoon $(Z_t)_{t\in\mathbb{Z}}$ iid satunnaismuuttujajono keskiarvolla nolla ja varianssilla yksi. ARCH-mallit (AutoRegressive Conditional Heteroskedasticity) ja näiden yleistyksenä GARCH-mallit (Generalized ARCH) ovat yleisesti muotoa

$$X_t = \sigma_t Z_t,$$

$t \in \mathbb{Z}$, missä prosessin *volatiliteetti* σ_t on mitallinen sigma-algebran $\mathcal{F}_{t-1} = \sigma(\{X_s | s \leq t-1\})$ suhteen, ja Z_t on riippumaton sigma-algebrasta $\mathcal{F}_{t-1}$. Näin määriteltyä sigma-algebraa $\mathcal{F}_t$ kutsutaan prosessin (X_t) historiaksi, ja se edustaa siis prosessista hetkeen t mennessä havaittua informaatiota; informaation kertymistä ajan suhteen kuvaa filtraatio $\mathbb{F} = (\mathcal{F}_t)_{t\in\mathbb{Z}}$. Mitallisuusoletuksen perusteella seuraa, että

$$\mathrm{Var}(X_t|\mathcal{F}_{t-1}) = \mathbb{E}(\sigma_t^2 Z_t^2|\mathcal{F}_{t-1}) = \sigma_t^2 \mathrm{Var}(Z_t) = \sigma_t^2.$$

Volatiliteetti σ_t on siis prosessin ehdollinen keskihajonta hetkellä t, ja muuttuu jatkuvasti prosessin aiempien (neliöityjen) arvojen funktiona; tästä nimitys "conditional heteroskedasticity".

Sarjan (X_t) sanotaan noudattavan ARCH(p)-prosessia, jos

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j}^2, \quad \forall\, t \in \mathbb{Z},$$

missä $\alpha_j > 0$, $j = 1, \ldots, p$. Vastaavasti (X_t) noudattaa GARCH(p, q)-prosessia, jos

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j}^2 + \sum_{k=1}^{q} \beta_k \sigma_{t-k}^2, \quad \forall\, t \in \mathbb{Z},$$

missä $\alpha_j > 0$, $j = 1, \ldots, p$ ja $\beta_k > 0$, $k = 1, \ldots, q$. Ehtona GARCH-prosessin (X_t) stationaarisuudelle ja äärelliselle varianssille on

$$\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k < 1.$$

ARCH-prosessin ideana on, että menneet havainnot vaikuttavat prosessin volatiliteettiin (ajan mukana vaimenevilla painoilla); näin ollen itseisarvoltaan suuret havainnot kasvattavat volatiliteettia hetkellisesti. GARCH-mallissa volatiliteetti hetkellä t riippuu paitsi menneistä havaintoarvoista, myös menneistä volatiliteetin arvoista. Tämä lisää volatiliteetin pysyvyyttä (volatility persistence) mallissa.

ARCH- ja GARCH-prosessit ovat periaatteessa autokorreloitumattomia WN-prosesseja, sillä niiden autokovarianssifunktiolle pätee

$$\begin{aligned}\gamma(h) &= \mathrm{Cov}(X_t, X_{t+h}) = \mathbb{E}\left((X_t - \mathbb{E}(X_t))(X_{t+h} - \mathbb{E}(X_{t+h}))\right) \\ &= \mathbb{E}(\sigma_t Z_t \sigma_{t+h} Z_{t+h}) = \mathbb{E}(Z_{t+h})\mathbb{E}(\sigma_t \sigma_{t+h} Z_t) = 0,\end{aligned}$$

kaikilla viiveillä h. Vaikka (X_t) on autokorreloitumaton prosessi, voidaan osoittaa että neliöity prosessi (X_t^2) (tai itseisarvoprosessi $(|X_t|)$) on voimakkaasti autokorreloitunut; itse asiassa prosessilla (X_t^2) on ARMA-tyyppinen rakenne. Ks. [5, luku 4.3].

4.3.2 ARMA-GARCH–mallin sovittaminen tuottoaikasarjaan

Finanssiaikasarjojen hallitseva piirre on volatiliteetin vaihtelu. Edellisen osion GARCH-malli vaikuttaa tässä suhteessa ominaisuuksiensa puolesta lupaavalta malliluokalta. Tuottoaikasarjat ilmentävät kuitenkin joskus myös autokorrelaatiota pienillä viiveillä h, ja mallin sopivuuden parantamiseksi on perusteltua lisätä edellisen kohdan GARCH-prosessiin $X_t = \sigma_t Z_t$ keskiarvotermi μ_t, jolloin malli saadaan muotoon

$$\begin{aligned} X_t &= \mu_t + \varepsilon_t, \\ \varepsilon_t &= \sigma_t Z_t, \end{aligned}$$

eli lyhyesti $X_t = \mu_t + \sigma_t Z_t$. Prosessin (μ_t) oletetaan noudattavan ARMA-mallia.

Edellä tehdyin oletuksin nähdään, että $\mathbb{E}(X_t|\mathcal{F}_{t-1}) = \mu_t$ ja $\mathbb{E}(X_t^2|\mathcal{F}_{t-1}) = \sigma_t^2$. Siis μ_t on X_t:n ehdollinen odotusarvo ja σ_t on ehdollinen keskihajonta, missä ehdollistus on prosessin havaitun historian suhteen.

Paperissa [36] tarkastellaan parsimoonista AR(1)-GARCH(1,1)–mallia prosessille (X_t), ja havaitaan tämän sopivan tuottoihin hyvin. Käytetään samaa mallispesifikaatiota. Ehdollinen odotusarvo μ_t noudattaa siis AR(1)-prosessia

$$\mu_t = \phi_0 + \phi_1 X_{t-1}, \tag{4.3}$$

ja (keskiarvolla adjustoidun sarjan $X_t - \mu_t = \varepsilon_t$) ehdollinen varianssi GARCH(1, 1)-prosessia

$$\sigma_t^2 = \alpha_0 + \alpha_1 (X_{t-1} - \mu_{t-1})^2 + \beta \sigma_{t-1}^2, \tag{4.4}$$

missä $\alpha_0, \alpha_1, \beta > 0$. Lisäksi vaaditaan, että $\alpha_1 + \beta < 1$ ja $|\phi_1| < 1$, jotta malli muodostaa stationaarisen prosessin äärellisellä varianssilla.

Tarkastellaan n havainnon mittaista (päivätuottojen) aikasarjaa, jolloin päivän t lopussa data koostuu havainnoista $(x_{t-n-1}, \ldots, x_{t-1}, x_t)$. Edellä määritelty

malli voidaan sovittaa tuottoaikasarjaan kvasi-suurimman uskottavuuden menetelmällä (QML).[7] Tässä tapauksessa maksimoidaan log-uskottavuusfunktio GARCH(1,1)-mallille t-jakautuneilla innovaatioilla parametriestimaattien saamiseksi. Vaikka lähestymistapa merkitsee että jakaumasta tehdään oletus, johon ei välttämättä uskota (siis innovaatioiden ei välttämättä uskota olevan t-jakautuneita), QML-menetelmän voidaan osoittaa tuottavan tarkentuvan ja asymptoottisesti normaalijakautuneen piste-estimaattorin (ks. [39]).

Mallin sovittamista varten innovaatioiden oletetaan siis noudattavan skaalattua t-jakaumaa, $Z_t \sim t(\nu, 0, (\nu-2)/\nu)$, missä skaalauksella saadaan jakauman varianssiksi yksi.[8] Maksimoitava uskottavuusfunktio on muotoa

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sigma_t} g\left(\frac{x_t - \mu_t}{\sigma_t}\right),$$

missä σ_t noudattaa GARCH(1,1)-mallia ja μ_t AR(1)-mallia, ja $g(z)$ on innovaatiojakauman tiheysfunktio.

Mallin sovituksen tuloksena saadaan parametriestimaatit $\hat{\boldsymbol{\theta}} = (\hat{\phi}_0, \hat{\phi}_1, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}, \hat{\nu})$. Estimaatit ehdollisen odotusarvon ja keskihajonnan sarjoille, $(\hat{\mu}_{t-n-1}, \ldots, \hat{\mu}_t)$ ja $(\hat{\sigma}_{t-n-1}, \ldots, \hat{\sigma}_t)$, saadaan rekursiivisesti malliyhtälöistä (4.3) ja (4.4) sopivien alkuarvojen valinnalla.[9] Näiden avulla voidaan määrittää standardoidut residuaalit $\boldsymbol{z}_t = (z_{t-n-1}, \ldots, z_t)$, missä

$$z_t = \frac{x_t - \hat{\mu}_t}{\hat{\sigma}_t}.$$

Mikäli malli sopii dataan, tulisi standardoitujen residuaalien muodostaa (likimain) iid sarja.

4.3.3 Innovaatiojakauman mallinnus GP-jakaumaa käyttäen

Innovaatiojakauman mallintamista varten pidetään edellisen osion mukaisia malliresiduaaleja iid realisaatioina innovaatiojakaumasta. Ylitemenetelmän mukaisesti tarkastellaan valitun korkean tason u ylittäviä tappioita, ja sovitetaan ylitteiden muodostamaan dataan yleistetty Pareto-jakauma,

$$G_{\xi,\beta}(z) = \begin{cases} 1 - \left(1 + \xi \frac{z}{\beta}\right)^{-\frac{1}{\xi}}, & \xi \neq 0, \\ 1 - e^{-z/\beta}, & \xi = 0, \end{cases}$$

missä $\beta > 0$; ks. kappaleet 1.5 ja 2.3 edellä.

[7] QML-menetelmää käytettäessä oletetaan, että aikasarjamallin dynaaminen muoto on spesifioitu oikein, mutta innovaatiojakauma voi olla spesifioitu väärin, tavanomaisesti olettaen innovaatioiden olevan normaalijakautuneita. Saatua uskottavuusfunktiota pidetään tällöin pikemminkin maksimoitavana tavoitefunktiona kuin aitona uskottavuutena.

[8] t-jakauman tiheysfunktio on $g(x; \nu, \mu, \sigma) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sigma\sqrt{\pi\nu}} \left(1 + \nu^{-1} \frac{(x-\mu)^2}{\sigma^2}\right)^{-\frac{\nu+1}{2}}$, missä $\nu > 0$ on vapausasteparametri ja $\Gamma(\cdot)$ gammafunktio.

[9] Voidaan esimerkiksi asettaa nämä yhtäsuuriksi otoskeskiarvon ja otoskeskihajonnan kanssa, tai usein myös yksinkertaisesti käyttää arvoina nollia.

Muistetaan (ks. esim. osio 2.3.3) että ylitejakauma voidaan kirjoittaa, kun $z \geq u$, seuraavasti:

$$\begin{aligned}\bar{G}(z) &= \mathbb{P}(Z > z) = \mathbb{P}(Z > u)\mathbb{P}(Z > z | Z > u) \\ &= \bar{G}(u)\mathbb{P}(Z - u > z - u | Z > u) = \bar{G}(u)\bar{G}_u(z - u).\end{aligned}$$

Kun ylitejakauma mallinnetaan GP-jakaumalla, $\bar{G}_u(z) = \bar{G}_{\xi,\beta}(z)$, saadaan tähän oletukseen perustuen edelleen

$$\bar{G}(z) = \bar{G}(u)\bar{G}_{\xi,\beta}(z - u) = \bar{G}(u)\left(1 + \xi\frac{z-u}{\beta}\right)^{-1/\xi}.$$

Yllä oleva lauseke antaa häntätodennäköisyydet, jos $\bar{G}(u)$ (tai siis $G(u)$) tunnetaan.

Oletetaan, että jakauman häntä alkaa kynnystasosta u; tällöin käytetystä n havainnon otoksesta satunnainen määrä N_u ylittää tämän tason. Estimoidaan $\bar{G}(u) = \mathbb{P}(Z > u)$ tuttuun tapaan näiden "häntähavaintojen" osuutena koko otoksesta, eli

$$\hat{\bar{G}}(u) = \frac{\hat{N}_u}{n} = \frac{k}{n},$$

kun on havaittu k tason u ylitystä – kuten aiemmin mainittu, suhteellinen osuus k/n on myös ylitystodennäköisyyden $\bar{G}(u)$ suurimman uskottavuuden estimaattori, jos ylityksien lukumäärä on Poisson- tai binomijakautunut. Häntäestimaattoriksi saadaan siis

$$\hat{\bar{G}}(z) = \frac{k}{n}\left(1 + \hat{\xi}\frac{z-u}{\hat{\beta}}\right)^{-1/\hat{\xi}}.$$

kun $z > u$. Käytännössä on mukavampaa *valita* k siten, että häntään jää tietty määrä havaintoja (siten, että $k << n$). Tämä menettely antaa satunnaisen kynnystason, vastaten havaintojen $(k+1)$. järjestystunnuslukua: Merkitään havaintojen järjestettyä otosta $z_{(1)} \geq z_{(2)} \geq \cdots \geq z_{(n)}$. Tällöin $u = z_{(k+1)}$, ja GP-jakauma sovitetaan ylitedataan $(y_1, \ldots, y_k) = (z_{(1)} - z_{(k+1)}, \ldots, z_{(k)} - z_{(k+1)})$, kuten aiemmin. Innovaatiojakauman häntäestimaattori voidaan nyt kirjoittaa muodossa

$$\hat{\bar{G}}(z) = \frac{k}{n}\left(1 + \hat{\xi}\frac{z - z_{(k+1)}}{\hat{\beta}}\right)^{-1/\hat{\xi}},$$

missä $\hat{\xi}$ ja $\hat{\beta}$ ovat sovitetun GP-jakauman parametrien SU-estimaatit. Kun $\alpha > 1 - k/n$, eli ollaan GP-jakaumalla mallinnetussa hännässä, α-kvantiilin z_α estimaattorille saadaan eksplisiittinen esitys kääntämällä yo. yhtälö:

$$\hat{z}_\alpha = \hat{z}_{\alpha,k} = z_{k+1} + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{1-\alpha}{k/n}\right)^{-\hat{\xi}} - 1\right).$$

4.3.4 Yhdistetty GARCH-EVT–malli riskimittojen estimoimiseen

Laitetaan edellä tarkastellut mallin palaset yhteen riskimittojen estimoimiseksi. Kuten edellä, olkoon F (negatiivisten) logaritmisten päivätuottojen muodostaman sarjan $(X_t)_{t\in\mathbb{Z}}$ reunajakauman kertymäfunktio, $X_t \sim F$. Muistetaan, että

prosessin (X_t) oletetaan noudattavan mallia

$$X_t = \mu_t + \sigma_t Z_t,$$

missä innovaatiot ovat WN-prosessi eli jono riippumattomia ja samoin jakautuneita satunnaismuuttujia yhteisellä kertymäfunktiolla G. Osion 4.1.1 mukaisesti X:n jakaumaa ja sen kvantiileja tarkastellessa voidaan erottaa kaksi tapausta: X:n stationaarisen jakauman kvantiili eli ehdollistamaton kvantiili on

$$x_\alpha = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\},$$

kun $0 < \alpha < 1$. Ehdollinen kvantiili puolestaan on ennustejakauman $F_{[X_{t+h}^{(h)}|\mathcal{F}_t]} = F_{[X_{t+1}+\cdots+X_{t+h}|\mathcal{F}_t]}$ kvantiili. Merkitään tätä

$$x_\alpha^t(h) = \inf\{x \in \mathbb{R} : F_{[X_{t+1}+\cdots+X_{t+h}|\mathcal{F}_t]} \geq \alpha\},$$

missä $h \in \mathbb{N}$ on tarkasteluhorisontti päivinä. $F_{[X_{t+h}^{(h)}|\mathcal{F}_t]}$ on siis h:n pituisen periodin tuottojen jakauma ehdollistettuna hetken t informaatiolla, eli hetkeen t mennessä havaitulla tuottohistorialla. Osiosta 4.2 muistetaan, että Value-at-Risk luottamustasolla α on suoraan tappiojakauman α-kvantiili. Siis $\mathrm{VaR}_\alpha^t(h) = x_\alpha^t(h)$.

Ei-ehdollistetulle Expected Shortfallille pätee jatkuvan jakauman tapauksessa $\mathrm{ES}_\alpha = \mathbb{E}(X|X > \mathrm{VaR}_\alpha)$, ja vastaava ehdollinen ES on

$$\begin{aligned}\mathrm{ES}_\alpha^t(h) &= \mathbb{E}\left(X_{t+h}^{(h)} | X_{t+h}^{(h)} > \mathrm{VaR}_\alpha^t(h), \mathcal{F}_t\right) \\ &= \mathbb{E}\left(\sum_{j=1}^h X_{t+j} | \sum_{j=1}^h X_{t+j} > \mathrm{VaR}_\alpha^t(h), \mathcal{F}_t\right).\end{aligned}$$

Keskitytään tarkastelemaan yhden askeleen riskimittoja ($h = 1$) ja merkitään näitä $\mathrm{VaR}_\alpha^t := \mathrm{VaR}_\alpha^t(1)$ ja $\mathrm{ES}_\alpha^t := \mathrm{ES}_\alpha^t(1)$. Nyt

$$\begin{aligned}F_{[X_{t+1}|\mathcal{F}_t]}(x) &= \mathbb{P}(X_{t+1} \leq x|\mathcal{F}_t) = \mathbb{P}(\mu_{t+1} + \sigma_{t+1}Z_{t+1} \leq x|\mathcal{F}_t) \\ &= \mathbb{P}\left(Z_{t+1} \leq \frac{x - \mu_{t+1}}{\sigma_{t+1}}|\mathcal{F}_t\right) = G\left(\frac{x - \mu_{t+1}}{\sigma_{t+1}}\right),\end{aligned}$$

koska $\mu_{t+1}, \sigma_{t+1} \in \mathcal{F}_t$ ja (Z_t) on iid prosessi. Ehdolliset riskimitat saadaan muotoon

$$\mathrm{VaR}_\alpha^t = x_\alpha^t = \mu_{t+1} + \sigma_{t+1}z_\alpha = \mu_{t+1} + \sigma_{t+1}\mathrm{VaR}_\alpha, \tag{4.5}$$

ja

$$\mathrm{ES}_\alpha^t = \mu_{t+1} + \sigma_{t+1}\mathbb{E}(Z|Z > z_\alpha) = \mu_{t+1} + \sigma_{t+1}\mathrm{ES}_\alpha, \tag{4.6}$$

missä $Z \sim G$ ja suureiden VaR_α, ES_α ymmärretään viittaavan jakaumaan G. Koska innovaatiojakauman hännän oletetaan noudattavan GP-jakaumaa, $G = G_{\xi,\beta}$, saadaan tuloksen (1.24) perusteella jakauman G Expected Shortfalliksi edelleen

$$\mathrm{ES}_\alpha = \frac{\mathrm{VaR}_\alpha}{1-\xi} + \frac{\beta - \xi u}{1-\xi}, \tag{4.7}$$

jolloin yhtälö (4.6) saadaan muotoon

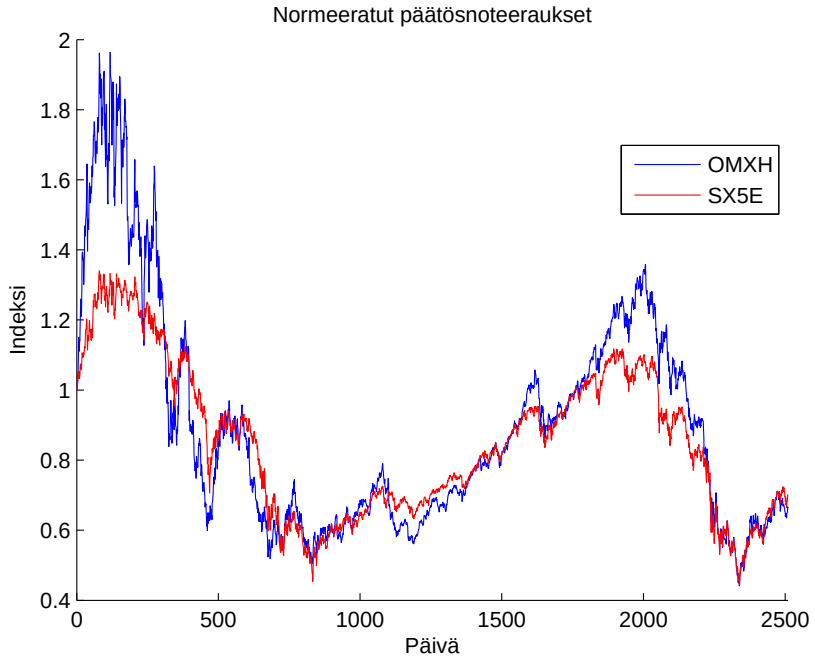
$$\mathrm{ES}_\alpha^t = \mu_{t+1} + \sigma_{t+1}\left(\frac{z_\alpha}{1-\xi} + \frac{\beta - \xi z_{(k+1)}}{1-\xi}\right), \tag{4.8}$$

missä $\xi < 1$ ja kynnystasona on $u = z_{(k+1)}$ edellisen osion mukaisesti.

4.4 Sovellus osakeindeksidataan

Sovelletaan edellä kehitettyä mallia osakemarkkinariskin mittaamiseen. Tarkasteltava data koostuu OMX Helsinki ja DJ Euro STOXX 50 –osakeindeksien (tickerit OMXH ja SX5E) päätöshinnoista aikaväliltä 10.11.1999 – 10.11.2009.[10] Nämä on edelleen muunnettu päivittäisiksi log-tuotoiksi. Esimerkki perustuu esitykseen [40].

Kuvassa 4.1 on näytetty OMXH- ja SX5E-indeksien kehitys tarkastelujaksolla, kun lähtöarvot on skaalattu ykköseksi. Kuvasta käy välittömästi ilmi osakemarkkinoille tyypillinen indeksien vahva korrelaatio. Muutoin nähdään, että OMXH nousi jakson alussa (ns. teknokuplan aikaan) selvästi (länsi)eurooppalaista osakemarkkinaa laajasti kuvaamaan pyrkivää SX5E:tä enemmän, mutta myös seurannut pudotus oli vastaavasti suurempi. Tämän jälkeen indeksien suhteelliset muutokset ovat vastanneet pitkälti toisiaan. Tarkastelujakson lopussa on nähtävissä finanssikriisin alkuun liittynyt raju pudotus molemmissa indekseissä.



Kuva 4.1: OMXH- ja SX5E-indeksien suhteellinen kehitys 10 vuoden tarkastelujaksolla.

Sovitetaan EVT-GARCH–malli tähän dataan. Data koostuu siis negatiivisista log-tuotoista eli tappioista $\boldsymbol{x}_t = (x_{t-n-1}, \ldots, x_t)$, missä kaupankäyntipäivien määrä $n = 2509$. Näihin sovitetaan osion 4.3.1 AR(1)-GARCH(1,1)-malli, jolloin saadaan estimaatit ehdollisille odotusarvoille $(\hat{\mu}_{t-n-1}, \ldots, \hat{\mu}_t)$ ja keskiha-

[10] Lähde: Bloomberg.

jonnoille $(\hat{\sigma}_{t-n-1}, \ldots, \hat{\sigma}_t)$, ja näiden avulla edelleen residuaaleille

$$\boldsymbol{z}_t = (z_{t-n-1}, \ldots, z_t) = \left(\frac{x_{t-n-1} - \hat{\mu}_{t-n-1}}{\hat{\sigma}_{t-n-1}}, \ldots, \frac{x_t - \hat{\mu}_t}{\hat{\sigma}_t} \right).$$

Hetki t vastaa havaintojakson viimeistä päivää, tässä tapauksessa päivämäärää 10.11.2009. Aikasarjamallin estimointitulokset on esitetty taulukossa 4.1. Parametriestimaatit poikkeavat tilastollisesti merkittävästi nollasta merkitsevyystasolla 0.05, OMXH-indeksille estimoidun mallin parametreja ϕ_1 ja α_0 lukuunottamatta. GARCH-parametrit β ovat molempien indeksien tapauksessa erittäin vahvasti merkitseviä.

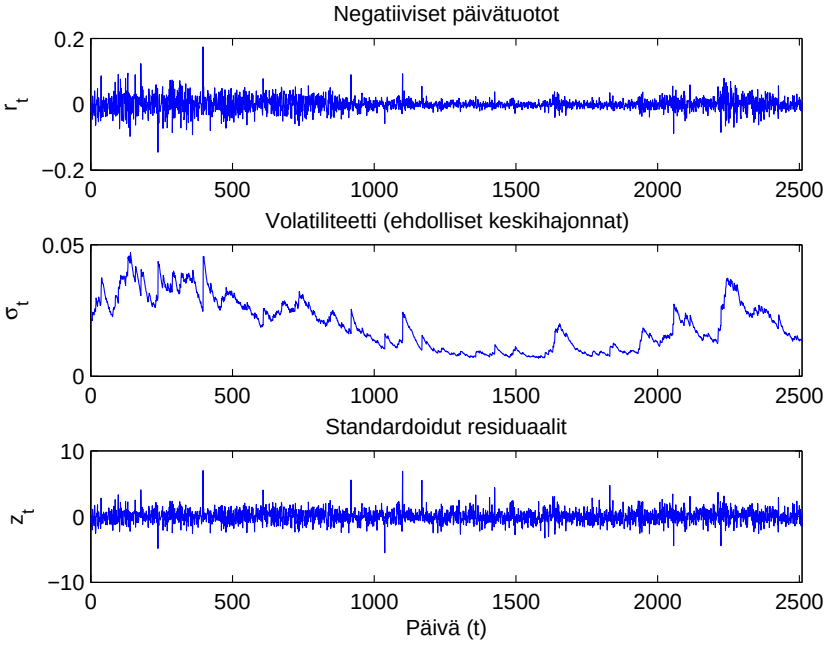
Taulukko 4.1: Tuottodataan sovitetun AR(1)-GARCH(1,1)-mallin parametriestimaatit.

Parametri	ϕ_0	ϕ_1	α_0	α_1	β	ν
OMXH						
SUE	0.000820	0.0227	$0.530 \cdot 10^{-6}$	0.0486	0.951	6.73
Keskivirhe	0.000252	0.0196	$0.318 \cdot 10^{-6}$	0.00723	0.00659	0.692
SX5E						
SUE	0.000532	-0.0493	$1.37 \cdot 10^{-6}$	0.0875	0.909	11.5
Keskivirhe	0.000217	0.0219	$0.493 \cdot 10^{-6}$	0.0108	0.0108	2.05

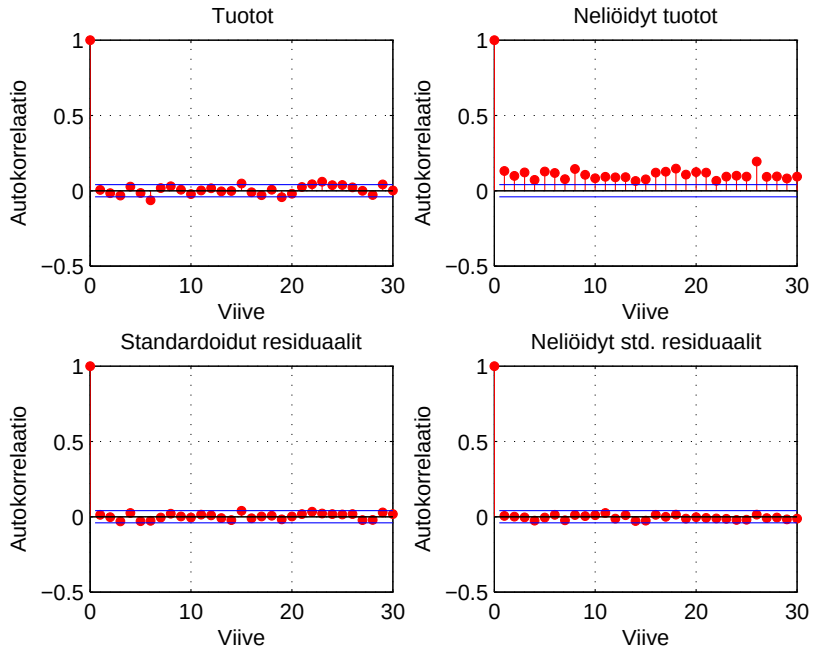
Tuottoaikasarja, estimoitu volatiliteetti ja standardoidut residuaalit on esitetty kuvassa 4.2 OMXH-indeksille; tilan säästämiseksi kuvat esitetään jatkossa vain OMXH:lle, sillä SX5E-indeksin vastaavat kuvaajat ovat laadullisilta ominaisuuksiltaan käytännössä täysin samanlaisia. Kuvasta nähdään selvästi, ettei tuottodata $\boldsymbol{x}_t = -\boldsymbol{r}_t$ ole iid:tä. Aikasarjassa voidaan havaita volatiliteetin muuttuminen ajan suhteen sekä volatiliteetin kasautuminen, vastaten "korkean aktiviteetin" periodeja. Kuvan alaosion standardoidut residuaalit sen sijaan vaikuttavat täyttävän iid-oletuksen kohtuudella.

Kuvaan 4.3 alla on piirretty autokorrelaatiofunktiot $\rho(h) = \gamma(h)/\gamma(0)$ tuotoille r_t ja standardoiduille residuaaleille z_t sekä näiden neliöidyille arvoille r_t^2 ja z_t^2. Tuottoaikasarjassa on havaittavissa mahdollisesti hieman autokorreloituneisuutta, mutta tämä on hyvin pientä. Sen sijaan neliöidyt tuotot ovat vahvasti autokorreloituneita; vrt. osion 4.1 tyylitellyt faktat luvun alussa. Standardoituja residuaaleja vastaavista kuvaajista kuvan alaosassa nähdään, että käytetty AR-GARCH–malli onnistuu poistamaan käytännössä kaiken autokorrelaation.

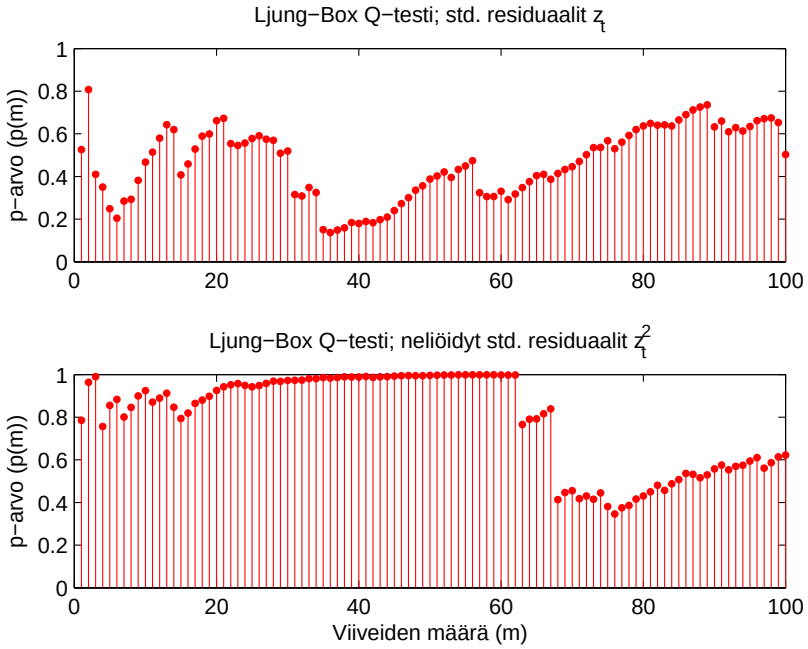
Autokorrelaatiofunktio tarjoaa havainnollisen tavan tutkia, ilmentääkö tarkasteltava aikasarja autokorrelaatiota tietyllä yksittäisellä viiveellä k. Testataan autokorreloituneisuutta vielä Ljung-Box-testiä käyttäen; tämä testaa autokorrelaation olemassaoloa useilla viiveillä samanaikaisesti. [41] Testin nollahypoteesina H_0 on, että ensimmäiset m autokorrelaatiota ovat kaikki nollia, $\rho(1) = \cdots = \rho(m) = 0$, ja tätä testataan vaihtoehtoista hypoteesia $H_1 : \exists\, k = 1, \ldots, m$ s.e. $\rho(k) \neq 0$ vastaan. Kuvaan 4.4 on piirretty Ljung-Box-testin p-arvot eri viiveiden määrillä m standardoiduille residuaaleille (yläosa) sekä näiden neliöidyille arvoille (alaosa). Testin perusteella nollahypoteesiä ei hylätä millään viiveiden lukumäärillä, ja johtopäätöksenä on, että residuaaleja voidaan pitää autokorreloitumattomina.



Kuva 4.2: Päivittäiset negatiiviset tuotot (x_t), estimoitu volatiliteetti ($\hat{\sigma}_t$) ja standardoidut residuaalit (z_t).



Kuva 4.3: Tuottojen ja neliöityjen tuottojen (ylärivi) sekä standardoitujen residuaalien ja näiden neliöiden (alarivi) autokorrelaatiofunktiot.



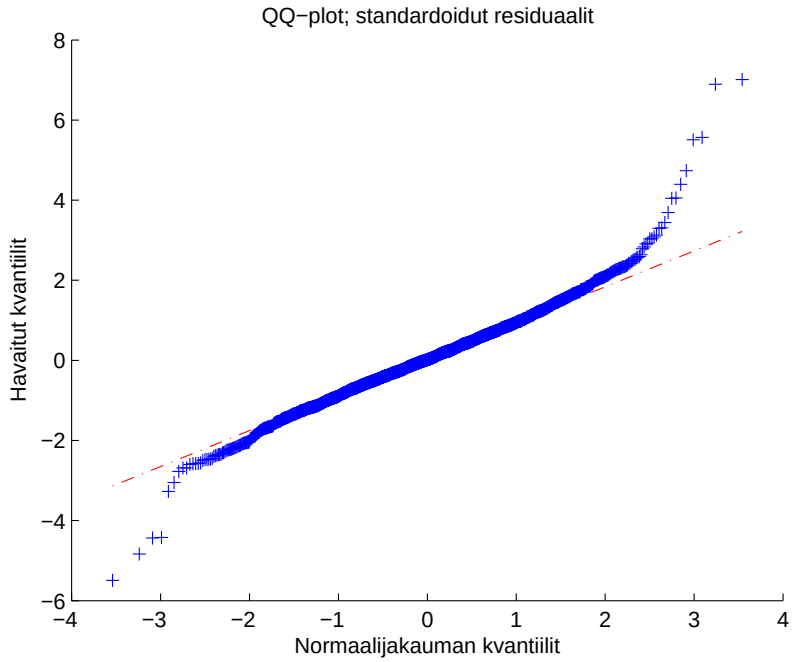
Kuva 4.4: Ljung-Box–testin p-arvot eri viivemäärillä m.

Esitettyjen tarkastelujen perusteella mallia voidaan pitää tilastollisessa mielessä riittävänä (adequate) kuvauksena tuottoaikasarjan ominaisuuksista.

Piirtämällä standardoitujen residuaalien empiirisen jakauman kvantiilit normaalijakauman kvantiileja vasten (kuva 4.5), havaitaan innovaatiojakauman olevan selvästi normaalijakaumaa paksuhäntäisempi. Lisäksi jakauman oikea häntä (vastaten tappioita) näyttää vasenta paksummalta, mikä viittaa siihen, ettei symmetrinen jakauma (kuten t-jakauma) välttämättä kuvaa tappio-/tuottojakaumaa erityisen hyvin.

Edellisessä osiossa keskityttiin tappiojakauman eli negatiivisten log-tuottojen jakauman oikean hännän tarkasteluun – täsmälleen samat tulokset pätevät tietysti vasemmalle hännällekin. Yleisemmin onkin tarpeen mallintaa tuotto- tai tappiojakauman molemmat hännät riskienhallintatarkoituksiin, sillä sijoitusinstrumentin hinnan nousu muodostaa riskin (tuottaa tappiota) sijoittajalle, jolla on lyhyt eli ns. shorttipositio kyseisessä instrumentissa. Esimerkki tällaisesta tilanteesta on vaikkapa osakkeen lyhyeksimyynti.

Mallinnetaan innovaatiojakauman molemmat hännät siis yleistettyä Pareto-jakaumaa käyttäen. Tätä varten valitaan kynnystasot u_U ja u_L siten, että molempiin häntiin jää 10 % äärihavainnoista (suurimmista ja pienimmistä). Tämä vastaa k:n arvoa 251, ja kynnyksiksi saadaan $u_U = z_{(k+1)} = 1.242\%$ ja $u_L = z_{(n-k)} = -1.163\%$ OMXH-indeksille. SX5E-indeksille kynnykset ovat $u_U = 1.361\%$ ja $u_L = -1.172\%$. Saadut parametriestimaatit on esitetty taulukossa 4.2.



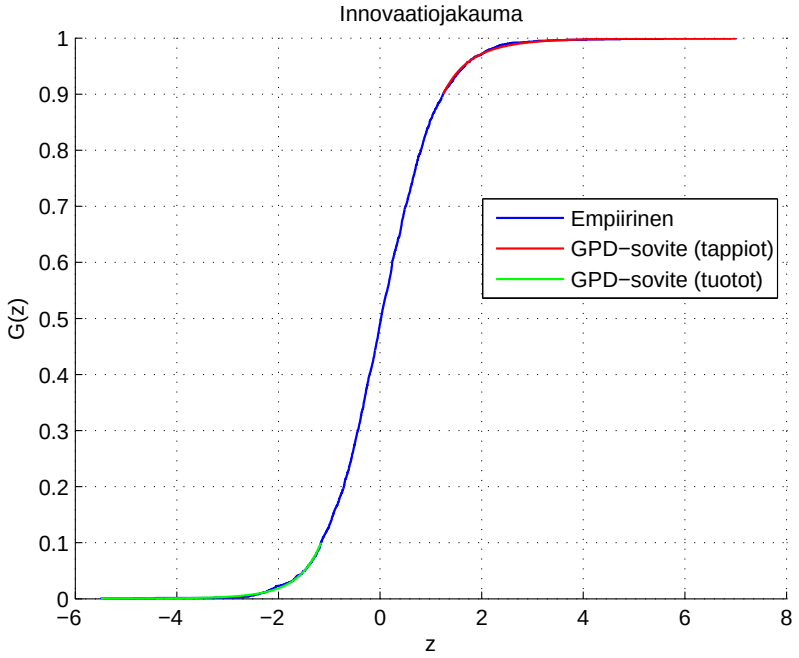
Kuva 4.5: Kvantiilikuvaaja; standardoidut residuaalit vs. normaalijakauma.

Taulukko 4.2: Parametriestimaatit innovaatiojakauman häntiin sovitetuille GP-jakaumille.

	Oikea häntä (tappiot)		**Vasen häntä** (tuotot)	
Parametri	SUE	95% luottamusväli	SUE	95% luottamusväli
OMXH				
ξ	0.147	$[0.021, 0.273]$	0.114	$[-0.025, 0.254]$
β	0.544	$[0.456, 0.648]$	0.454	$[0.377, 0.548]$
SX5E				
ξ	0.0193	$[-0.0846, 0.124]$	-0.0803	$[-0.179, 0.0178]$
β	0.546	$[0.464, 0.642]$	0.500	$[0.427, 0.586]$

Taulukosta havaitaan, että SX5E-indeksin tappiojakauman vasemman hännän – tuottojakauman oikean hännän – muotoparametrin estimaatti on negatiivinen; tämä viittaa siihen, että tuottojakaumalla olisi äärellinen oikea päätepiste.[11] Ei ole kuitenkaan mitään syytä, miksi tuottojen oletettaisiin olevan rajoitettuja. Käytännön sovelluksissa voikin olla syytä asettaa tällainen negatiivinen arvo nollaksi, ainakin mikäli voittojen äärikvantiilien mallintaminen on tärkeää käsitellyn sovelluksen näkökulmasta; vrt. kuitenkin alla SX5E:n tappiojakauman vasemmalle hännälle saatuihin riskilukuihin.

Sovitetut GPD-hännät yhdessä empiirisen kertymäfunktion kanssa on piirretty kuvaan 4.6. Päivätuottohavaintoja on sen verran paljon, että empiirinen jakauma vaikuttaa varsin sileältä jakauman keskivaiheilla.



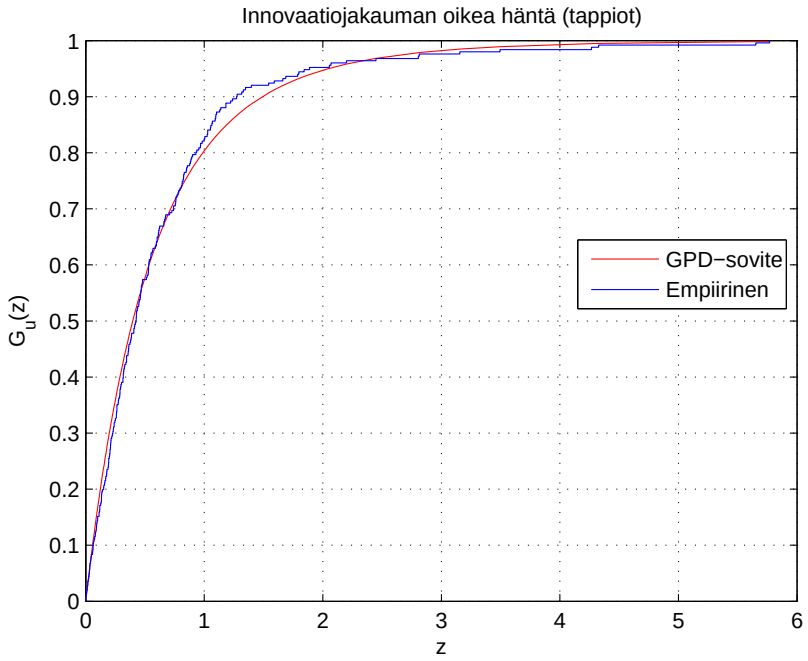
Kuva 4.6: Empiirinen innovaatiojakauma sovitetuilla GPD-hännillä.

Kuvissa 4.7 ja 4.8 on esitetty tarkemmin jakauman hännät. Nähdään, että GP-jakaumat sopivat häntähavaintoihin varsin hyvin.

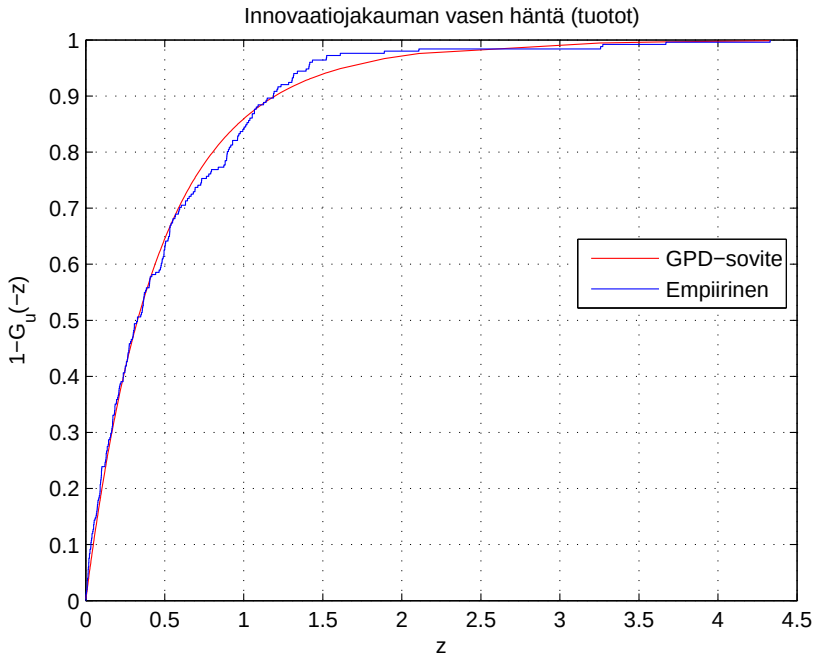
Käytetään sovitettua mallia ehdollisen Value-at-Riskin ja Expected Shortfallin estimoimiseen osakeindekseille. Riskimittojen arvot on esitetty taulukossa 4.3 tappioille ja taulukossa 4.4 voitoille. Nämä tulee siis tulkita hetken t estimaateiksi (hetkeen t mennessä kertyneeseen informaatioon perustuen) seuraavan päivän, $t+1$, markkinariskistä.

On mielenkiintoista huomata, että (hajautetumpaan) SX5E-indeksiin liittyvä tappioriski – tarkastelluilla riskimitoilla mitattuna – on pienemmillä α:n arvoil-

[11] Äärellinen (oikea) päätepiste, $x_F < \infty$, tarkoittaa että jakauma on *lyhythäntäinen*. Tämä ei kuitenkaan välttämättä merkitse, että jakauma olisi *kevythäntäinen*.



Kuva 4.7: Innovaatiojakauman oikea häntä, vastaten tappioita.



Kuva 4.8: Innovaatiojakauman vasen häntä, vastaten voittoja.

Taulukko 4.3: Estimoidut 1-päivän riskimittojen arvot; tappiot (oikea häntä).

Luottamustaso	**0.90**	**0.95**	**0.99**	**0.995**	**0.999**	**0.9997**
OMXH, VaR_α^t	1.69 %	2.23 %	3.71 %	4.46 %	6.54 %	8.45 %
OMXH, ES_α^t	2.56 %	3.19 %	4.92 %	5.80 %	8.24 %	10.5 %
SX5E, VaR_α^t	2.03 %	2.61 %	3.96 %	4.56 %	5.98 %	7.07 %
SX5E, ES_α^t	2.87 %	3.45 %	4.83 %	5.44 %	6.89 %	8.00 %

Taulukko 4.4: Estimoidut 1-päivän riskimittojen arvot; voitot (vasen häntä).

Luottamustaso	**0.90**	**0.95**	**0.99**	**0.995**	**0.999**	**0.9997**
OMXH, VaR_α^t	1.59 %	2.03 %	3.21 %	3.78 %	5.31 %	6.66 %
OMXH, ES_α^t	2.43 %	3.06 %	4.75 %	5.62 %	7.98 %	10.2 %
SX5E, VaR_α^t	1.75 %	2.26 %	3.33 %	3.75 %	4.64 %	5.23 %
SX5E, ES_α^t	2.84 %	3.36 %	4.65 %	5.22 %	6.59 %	7.65 %

la suurempi kuin "periferisempään" OMXH-indeksiin liittyvä. Toisaalta suurilla luottamustasoilla johtopäätös kääntyy päinvastaiseksi, niin kuin ennakolta saattaisi olettaakin. Vertailussa indeksien välillä täytyy kuitenkin huomata, että tarkastellut riskimitat ovat ehdollisia: kummankin indeksin estimoituun riskiin hetkellä t vaikuttavat siis niitä koskevat markkinaolosuhteet estimoidun volatiliteetin kautta.

4.4.1 Pidemmän horisontin riskin simulointi

Tarkastellaan lyhyesti riskimittojen estimoimista usean päivän tuottoperiodille, eli tapauksessa $h > 1$. GARCH-malleille jakauma $F_{[X_{t+1}+\cdots+X_{t+h}|\mathcal{F}_t]}$ ei ole tiedossa analyyttisesti edes tunnetun innovaatiojakauman tapauksessa ([36]), joten riskimittojen estimointi täytyy perustaa simulointilähestymistapaan.

Merkitään negatiivisiin log-tuottoihin perustuvan innovaatiojakauman oikeaan häntään (tappioihin) sovitetun GP-jakauman parametriestimaatteja $\hat{\xi}^U$, $\hat{\beta}^U$, ja vasempaan häntään (voittoihin) liittyviä estimaatteja vastaavasti $\hat{\xi}^L$, $\hat{\beta}^L$. Mallinnetaan innovaatiojakauman hännät GP-jakaumilla, ja keskiosa ei-parametrisesti empiiristä jakaumaa käyttäen (vrt. kuva 4.6). Näin muodostetusta jakaumasta G voidaan simuloida iid innovaatioita $Z \sim G$ käyttämällä bootstrap-menetelmän ja GPD-simulaation yhdistelmää seuraavasti [36, 42]:

Algoritmi 4.1 (Satunnaislukujen generointi innovaatiojakaumasta)

1. Poimitaan satunnaisesti yksi residuaali AR-GARCH-mallin sovituksessa saadusta n:n residuaalin otoksesta; olkoon tämä z_0.
2. Jos $z_0 > u^U = z_{(k+1)}$, generoidaan $G_{\hat{\xi}^U,\hat{\beta}^U}$-jakautunut satunnaismuuttuja y^U GP-jakaumasta kynnyksellä 0 (ks. algoritmi 3.3, kohta 2), ja palautetaan $z_{(k+1)} + y^U$;

3. Jos $z_0 < u^L = z_{(n-k)}$, generoidaan $G_{\hat{\xi}^L,\hat{\beta}^L}$-jakautunut satunnaismuuttuja y^L GP-jakaumasta kynnyksellä 0, ja palautetaan $z_{(n-k)} - y^L$;
4. Muutoin (kun $z_{(n-k)} \leq z_0 \leq z_{(k+1)}$) palautetaan residuaali z_0 itse.
5. Toistetaan menettely.

Simulointialgoritmi generoi siis pisteitä jakaumasta

$$\hat{G(z)} = \begin{cases} 1 - \frac{k}{n}\left(1 + \hat{\xi}^U \frac{z - z_{(k+1)}}{\hat{\beta}^U}\right), & z > z_{(k+1)}, \\ \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{z_i \leq z\}}, & z_{(n-k)} \leq z_0 \leq z_{(k+1)}, \\ \frac{k}{n}\left(1 + \hat{\xi}^L \frac{|z - z_{(n-k)}|}{\hat{\beta}^L}\right), & z < z_{(n-k)}. \end{cases}$$

Käyttäen innovaatiojakaumaa $\hat{G(z)}$ ja sovitettua AR(1)-GARCH(1,1)-mallia voidaan simuloida (negatiiviset) tuotot $(x_{t+1}, \ldots, x_{t+h})$. Näiden kumulatiivinen summa $x_{t+h}^{(h)} = \sum_{j=1}^{h} x_{t+j}$ on yksi realisaatio h-periodin tuotosta $X_{t+h}^{(h)}$. Toistamalla simulaatio riittävän monta kertaa saadaan estimaatti tuottojen jakaumalle eli tuottojakaumalle $F_{[X_{t+h}^{(h)}|\mathcal{F}_t]} = F_{[X_{t+1}+\cdots+X_{t+h}|\mathcal{F}_t]}$, ja riskimitat voidaan estimoida suoraviivaisesti tähän perustuen, kuten aiemmin.

Tarkastellaan esimerkkinä 1-vuoden VaR:in ja ES:n estimointia. Tämä vastaa siis tarkasteluhorisonttia $h = 250$, kun vuodessa oletetaan olevan keskimäärin 250 kaupankäyntipäivää. Siirrytään käyttämään aritmeettisia tuottoja, $r_t^A = S_t/S_{t-1} - 1$, logaritmisten tuottojen $r_t^L = \ln(S_t/S_{t-1})$ sijasta, koska pitkän aikavälin simulaatiossa logaritmisten tuottojen käyttö voi aiheuttaa ongelmia (esim. kokonaistuotto eli yksittäisten log-tuottojen summa voi mennä alle -100 %:n). Logaritmiset tuotot saadaan muutettua aritmeettisiksi käyttämällä relaatiota $r_t^A = \exp(r_t^L) - 1$.

Taulukossa 4.5 on esitetty estimoidut VaR- ja ES-luvut OMXH-indeksille 10^6 simulaatiopolkuun perustuen. Kussakin simulaatiossa simuloitiin siis tuotot päivä kerrallaan, ja näistä rakennettiin simulaatiopolun kokonaistuotto hetkellä $t + h$ eli vuoden päästä.

Taulukko 4.5: 1-vuoden riskimittojen estimaatit OMXH-indeksille hetken t = 10.11.2009 tilanteesta.

Luottamustaso	**0.90**	**0.95**	**0.99**	**0.995**	**0.999**	**0.9997**
VaR_α^t	28.7 %	34.6 %	46.8 %	51.8 %	63.3 %	69.5 %
ES_α^t	36.8 %	42.1 %	53.8 %	58.6 %	69.0 %	75.7 %

4.4.1.1 Vertailu Solvenssi II:een

Kuten aiemmin on mainittu, esimerkiksi Solvenssi II –vakavaraisuussäännöstön kvantitatiivisissa vaatimuksissa käytetään pääomavaateiden laskemisessa 1-vuoden 99.5 % Value-at-Riskiä vastaavia "shokkeja", eli tämän luvun terminologiaa käyttäen mittaa $\mathrm{VaR}_{0.995}^t(250)$ (missä t on periaatteessa raportointipäivä,

eli vuoden tai kvartaalin viimeinen päivä). Yllä olevan taulukon $\mathrm{VaR}^t_{0.995}$ voidaan siis tulkita tarkastellun – SII-terminologiassa (osittaisen) sisäisen – mallin antamaksi osakeshokin arvoksi suomalaisille osakkeille (laskentahetken ollessa vuoden 2009 loppupuoli). Arvoksi saadaan n. 52 %.

Solvenssi II:ta vastaavassa Sveitsissä käytössä olevassa Swiss Solvency Testissä (SST) käytetty 1-vuoden Expected Shortfall tasolla 99 % eli tässä $\mathrm{ES}^t_{0.99}(250)$ antaa osakeshokin arvoksi n. 54 %. Tämä on tässä tapauksessa vain hieman suurempi kuin SII:ssa käytetylle riskimitalle mallissa saatu 52 %.

Vertailun vuoksi viidennessä Solvenssi II:sen vaikuttavuusarvioinnissa eli QIS5-spesifikaatiossa osakeshokkina käytettiin globaaleille osakkeille arvoa 30 % ja muille osakkeille arvoa 40 %; nämä sisälsivät markkinoiden tilanteen huomioimaan pyrkivän ns. Symmetric Adjustment (SA) –tekijän -9 %, perusshokkien ollessa 39 % ja 49 %. Kirjoitushetkellä viimeisimmissä, uudistetuissa teknisissä spesifikaatioissa ("Revised Technical Specifications for the Solvency II valuation and Solvency Capital Requirements calculations") luokittelu muutettiin tyypin 1 (ETA- tai OECD-maissa säännellyillä markkinoilla listatut) ja tyypin 2 (muut) osakeriskillisiin arvopapereihin ja SA:n arvo vahvistettiin -7 %:ksi, mutta muuten osakeriskin laskenta säilyi aiemmanlaisena.[12] Osakeshokit olivat siis 32 % ja 42 %. Nämä shokit koskevat Solvenssi II:n standardikaavan käyttöä.

Tässä osiossa saatuja tuloksia ei voi aivan suoraan verrata Solvenssi II:n spesifikaatioiden käyttämiin arvoihin siinä mielessä, että osakeriski on estimoitu vuoden 2009 loppupuolen tilanteesta, kun QIS5-spesifikaatio on julkaistu heinäkuussa 2010 ja uudistetut spesifikaatiot joulukuussa 2012. Toisaalta SII:n spesifikaatiot ovat pysyneet useita vuosia samoina. Voidaan todeta, että suuremman perusshokin arvo 49 % on lähellä tämän luvun mallin suomalaisille osakkeille antamaa arvoa. Sen sijaan ETA- tai OECD-maihin, joihin Suomikin kuuluu, sovellettava 39 % on selvästi pienempi. Täytyy muistaa, että SII:n tavoitteena on kuitenkin asettaa riski tasolle, joka vastaa odotusarvoisesti kerran 200 vuodessa ylitettävää (vuositason) tappiota.

Spesifikaatioiden ns. symmetrisen säätötekijan (SA) huomiointi alentaa shokkien arvoja entisestään. Tämän säätötekijän tai "heilunnan vähentäjän" (equity dampener) tarkoituksena on vähentää säätelyn prosyklisiä vaikutuksia, siten ettei pääomavaatimuksen kasvu esim. lamassa johtaisi tilanteeseen, jossa vakuutusyhtiö joutuu realisoimaan osakeomistuksiaan tappiolla lyhyen aikavälin sääntelyvaatimusten täyttämiseksi. SA-tekijän arvo määritetään vertaamalla laskentahetken MSCI World -indeksin arvoa ko. indeksin 3-vuoden liukuvaan keskiarvoon. Mikäli osakkeiden hinnat ovat keskiarvoaan korkeammalla, kasvatetaan osakeshokin arvoa, ja mikäli hinnat ovat liukuvaa keskiarvoa alempana, vähennetään shokkia. Säätötekijän tavoite, sääntelyn aiheuttaman prosyklisyyden vähentäminen, voidaan nähdä sinänsä perusteltuna ja kannatet-

[12] Uudistettuihin teknisiin spesifikaatioihin perustuvassa ns. LTGA-arviointipaketissa sen sijaan käytettiin "siirtymävaihetta" (ns. transitional measure) Solvenssi I:sen laskentatavasta Solvenssi II:sen laskentatapaan, siten kuin julkaistut tekniset spesifikaatiot jälkimmäistä tällä hetkellä edustavat. Siirtymä tarkoitti mm., että SI:n mukaisesta diskonttokäyrästä siirrytään vähitellen SII:n mukaiseen projektion kuluesssa. Osakeriskin osalta LTGA-spesifikaatioissa säädetään, että "siirtymä" huomioidaan siten, että SA-tekijää ei käytetä, ja osakeshokin arvona käytetään 22 %:ia kaikille osakkeille. Sekä QIS- että LTGA-spesifikaatiot löytyvät EIOPA:n verkkosivuilta, osoitteesta `https://eiopa.europa.eu/consultations/qis/insurance/index.html`.

tavana. Samalla täytyy kuitenkin todeta, että määrittely nykymuodossaan on luonteeltaan *ad hoc*, ja sotii Solvenssi II:n kvantitatiivisten pääomavaatimusten perustana olevaa markkinaehtoista lähestymistapaa vastaan. Symmetric Adjustmentin kaltaisten "ylimääräisten" säätötekijöiden käyttöönoton taustalla on myös nähty poliittinen vaikutus, siten että tämänkaltaisten keinojen sopivalla käytöllä varmistetaan, että tuloksena saadaan pääomavaatimus jonka kanssa vakuutusyhtiöt voivat elää.[13] Kaikkien osapuolten kannalta läpinäkyvämpää olisi pyrkiä kuvaamaan riskiä mahdollisimman todellisuuden mukaisesti sellaisena kuin se on, vaikka tämä tuottaisi epämukavan suuria pääomavaatimuksia – sekä yhtiöiden että valvojien kannalta – jolloin keskustelu kääntyy olennaiseen eli siihen, minkäsuuruiseen riskiin lainsäätäjät ja valvojat velvoittavat vakuutusyhtiöt varautumaan pääomia varaamalla, ja miten tämä vaikuttaa yhtiöiden riskinottokykyyn sekä vakuutustoiminnan kehitykseen laajemmin.

[13]Ks. esim. kommentti `http://www.barrhibb.com/blog/entry/a_comment_on_the_solvency_ii_equity_dampener/`.

Luku 5

Yhteenveto

Tässä tutkimuksessa tarkasteltiin ääriarvoteorian soveltamista reaalimaailman ongelmiin. Tavoitteena esityksellä oli luoda katsaus klassiseen ääriarvoteoriaan ja kentän vakiintuneeksi muodostuneiseen nykymenetelmiin. Matemaattista taustateoriaa käsiteltiin vain sen verran kuin nähtiin tarpeelliseksi käytettyjen menetelmien ymmärtämiseksi tai perustelemiseksi. Sen sijaan painotus esityksessä on tilastollisilla menetelmillä ja näiden soveltamisella.

Esityksessä tarkasteltiin kolmea sovellusta: merenpinnan korkeuden mallintamista, onnettomuuskuolemien määrän arviointia ja markkinariskin mallinnusta.

Vedenkorkeuden mallintamisongelma toimi luvussa 2 viitekehyksenä, jonka avulla esiteltiin esityksessä käsitellyt ääriarvoteoriaan perustuvat tilastolliset menetelmät. Standardien blokkimaksimi- ja ylitemenetelmän lisäksi vedenkorkeutta mallinnettiin mm. epästationaarisella GEV-mallilla ja epästationaarisilla pisteprosessimalleilla. Trendin sisällyttäminen malleihin paransi mallin sopivuutta huomattavasti, ja tarkastelun perusteella vedenkorkeusmaksimeissa eli vedenkorkeuden ääriarvoissa on havaittavissa kasvava trendi. Tällaisten ominaisuuksien huomioiminen on keskeistä ennustettaessa ilmiön käytöstä tulevaisuuteen. Trendin sisältäviä pisteprosessimalleja käytettiin alimman tulvien kannalta hyväksytyn rakentamiskorkeuden arviointiin Helsingin rannikolla. Mallinnusta laajennettiin myös ottamalla pisteprosessimalliin mukaan selittävänä muuttujana ilmanpaine-eroon perustuva North Atlantic Oscillation (NAO) –indeksi. NAO-indeksin lisääminen paransi mallia merkittävästi trendin vaikutuksen huomioimisen jälkeenkin.

Luvussa 3 mallinnettiin suomalaisten onnettomuuskuolemien määrää ääriarvoteoriaa käyttäen. Koska suurista onnettomuuksista on vain vähän havaintodataa, tarkasteltiin myös Ruotsia koskevan onnettomuuskuolemadatan yhdistämistä Suomea koskevaan ja tilastollista estimointia yhdistettyyn aineistoon perustuen. Laajennettuun aineistoon sovitettu malli osoittautui kuvaavan suomalaisten onnettomuuskuolemamääriä selvästi pelkkään Suomea koskevaan aineistoon perustuvaa mallia paremmin. Trendinomaisten piirteiden olemassaoloa onnettomuuskuolemadatassa tutkittiin myös mallintamalla kuolemien määriä epästationaarisilla pisteprosesseilla. Johtopäätöksenä oli, ettei datassa ole viit-

teitä trendistä, kun tarkastellaan sovelluksessa kiinnostuksen kohteena olleita suuria onnettomuuksia.

Luvussa tarkasteltiin myös katastrofiriskin mittaamista estimoitua mallia käyttäen ja esitettiin menetelmät onnettomuuskuolemien määrien sekä niiden otosmaksimien simuloimiseksi. Onnettomuuskuolemamallin antamia tuloksia verrattiin QIS5- ja LTGA-vaikuttavuusarvioinneissa käytettyihin katastrofikuolemia koskeviin skenaarioihin. Johtopäätöksenä on, että kummankin spesifikaation käyttämä massaonnettomuusskenaario vaikuttaa tarkastellun mallin perusteella selvästi Solvenssi II:ssa tavoitteena olevaa 1-vuoden 99.5 %:n Value-at-Risk –tasoa riskillisemmältä. LTGA:n joukko-onnettomuusskenaarion implikoima onnettomuuskuolemien määrä on 1 890 henkeä, kun onnettomuuskuolemien estimoitu 99.5 %:n VaR oli 410 henkeä – ja 1 890 henkeä vastaa EVT-mallissa 1-vuoden 99.86 %:n VaR:ia. QIS 5:n areenariskisenaarion implikoima riskitaso on vielä tätä suurempi ollen 2 500 henkeä.

Luvussa 4 tarkasteltiin finanssiaikasarjojen ääriarvojen ja erityisesti markkinariskin mallinnusta. Finanssiaikasarjat ilmentävät tyypillisesti epätriviaalia riippuvuusrakennetta, ja niille on ominaista matalan ja korkean aktiviteetin jaksojen vaihtelu sekä itseisarvoltaan suurien arvojen (tappioiden ja tuottojen) kasaantuminen ajassa: ts. tuottoprosesseja kuvaavia ominaisuuksia ovat stokastinen volatiliteetti sekä volatiliteetin kasaantuminen ja pysyvyys. Tämän riippuvuusrakenteen huomioiminen on välttämätöntä markkinariskin mallintamiseksi realistisesti. Luvussa esitetyssä lähestymistavassa tuottodataan sovitettiin ensin GARCH-tyypin malli volatiliteettirakenteen mallintamiseksi, ja sovituksen tuloksena saadun residuaalijakauman hännät mallinnettiin ääriarvoteoriaa hyödyntäen. Rakennettu malli osoittautui kuvaavan tuottoprosessia varsin hyvin esitettyjen testien perusteella. Malliin perustuen esitettiin menetelmät riskiä kuvaavien riskimittojen (Value-at-Risk, Expected Shortfall) estimoimiseksi, sekä pidemmän horisontin riskin mittaamiseksi simulointilähestymistapaa käyttäen.

Kuten edellä esitetyistä sovelluksistakin havaitaan, ääriarvoteorian sovellusalue on varsin laaja, ja se tarjoaa perustan monien hyvin erilaisten ääri-ilmiöiden kuvaamiseen. Sellaisenaan ääriarvoteorian tarjoamien soveltavien menetelmien tunteminen on välittömästi hyödyllistä kaikille kvantitatiivisen riskienhallinnan parissa työskenteleville. Ääri-ilmiöt ovat jo määritelmänsä mukaan harvinaisia, mikä tarkoittaa että niistä on usein olemassa vain niukasti havaintoja. Näin ollen kaikkeen havaintoihin perustuvaan tilastolliseen mallinnukseen liittyy tässä yhteydessä suuri epävarmuus. Ääriarvoteoria tarjoaa kuitenkin perustellun, ja täsmällisen matemaattisen teorian tukeman, lähestymistavan harvinaisten ilmiöiden kuvaamiseen ja ilmiöihin liittyvän epävarmuuden arviointiin.

Ytimekkäänä yhteenvedonomaisena lopetuksena voitaneen esittää seuraava Jonathan Tawnilta peräisin oleva sitaatti (tämäkin on otettu teoksesta [2]): *"The key message is that EVT cannot do magic — but it can do a whole lot of better than empirical curve-fitting and guesswork. My answer to sceptics is that if people arent given well-founded statistical methods like EVT, they just use dubious ones instead."*

Liite A

Suurimman uskottavuuden menetelmä

Tarkastellaan riippumattomien ja samoin jakautuneiden satunnaismuuttujien jonoa $X_1, X_2, \ldots$ ja oletetaan, että näiden yhteinen tiheysfunktio f on olemassa. Olkoon havaittu otos $\boldsymbol{X} = \boldsymbol{x} = (x_1, \ldots, x_n)$, joka siis on satunnaismuuttujien $X_1, \ldots, X_n$ realisaatio. Suurimman uskottavuuden menetelmää sovellettaessa valitaan malliperhe, jonka piiriin oletetaan tarkasteltavien satunnaismuuttujien jakauman kuuluvan, ja estimoidaan jakauman tuntemattomat parametrit; kyse on siis parametrisesta estimoinnista. Olkoon todennäköisyysjakaumien perhe $\mathscr{F} = \{f(\boldsymbol{x}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, missä $\boldsymbol{\theta}$ parametrivektori (tai skalaari) ja Θ parametriavaruus. Esimerkiksi Poisson-jakauman tapauksessa tuntemattomia parametreja on yksi, $\theta = \lambda$, ja normaalijakauman tapauksessa kaksi, $\boldsymbol{\theta} = (\mu, \sigma)$.

Jokainen parametrivektorin arvo $\boldsymbol{\theta} \in \Theta$ määrittelee mallin $\mathcal{M} \in \mathscr{F}$ joka liittää (mahdollisesti) eri todennäköisyydet havaittuun dataan. Havaitun datan todennäköisyyttä parametrien funktiona kutsutaan uskottavuusfunktioksi (likelihood function), merkitään L. Uskottavuusfunktio on siis yksinkertaisesti tiheysfunktio (tai todennäköisyysfunktio diskreeteillä jakaumilla) tulkittuna parametrien $\boldsymbol{\theta}$ funktioksi kiinnitetylle (eli havaitulle) datalle $\boldsymbol{x}$. Täsmällisemmin, iid datalle

$$L(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}). \tag{A.1}$$

Intuitiivisesti, parametrien arvot $\boldsymbol{\theta}$, jotka liittävät korkean todennäköisyyden havaittuun dataan, ovat uskottavampia kuin arvot, jotka liittävät matalamman todennäköisyyden havaittuun otokseen. Suurimman uskottavuuden estimoinnissa periaatteena on valita malli, jolla on suurin uskottavuus, sillä kaikista tarkasteltavista vaihtoehtoisista malleista $\mathcal{M} \in \mathscr{F}$ juuri sen mukaan on todennäköisintä havaita sellainen otos kuin havaittiin.

Parametrivektorin $\boldsymbol{\theta}$ suurimman uskottavuuden estimaatti (SU-estimaatti, SUE) $\hat{\boldsymbol{\theta}}$ saadaan siis maksimoimalla uskottavuusfunktion L arvo parametrien $\boldsymbol{\theta}$ suhteen:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \boldsymbol{x}).$$

Jatkossa merkitään myös $\hat{\boldsymbol{\theta}}_n$, kun halutaan korostaa estimaattorin riippuvuutta otoskoosta n.

Yleensä on mukavampi muuttaa termien tulo yhtälössä (A.1) summaksi ottamalla kaavassa logaritmi puolittain, ja tarkastella logaritmista uskottavuutta eli log-uskottavuutta

$$l(\boldsymbol{x};\boldsymbol{\theta}) = \ln L(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i;\boldsymbol{\theta}). \tag{A.2}$$

Logaritmifunktio $l(\cdot)$ on monotoninen, joten funktioilla L ja l on samat maksimikohdat, eikä muunnos siten vaikuta saataviin estimaatteihin $\hat{\boldsymbol{\theta}}$.

Kun satunnaismuuttujat $(X_i)_{i=1}^n$ ovat riippumattomia mutta eivät samoin jakautuneita, uskottavuusfunktioksi tulee suoraan

$$L(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(x_i;\boldsymbol{\theta}),$$

log-uskottavuudeksi vastaavasti

$$l(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f_i(x_i;\boldsymbol{\theta}),$$

missä f_i on X_i:n tiheysfunktio (tai todennäköisyysfunktio). Yleisemmin, kun $\mathscr{F} = \{f(\boldsymbol{x};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ on perhe yhteisjakaumien tiheysfunktioita satunnaismuuttujavektorille $X = (X_1, \ldots, X_n)$ joka ei välttämättä ole riippumaton, havaittuun otokseen $\boldsymbol{x} = (x_1, \ldots, x_n)$ perustuva uskottavuus on

$$L(\boldsymbol{x};\boldsymbol{\theta}) = f(\boldsymbol{\theta};\boldsymbol{x}).$$

A.1 Suurimman uskottavuuden estimaattorin ominaisuuksia

Tiettyjen säännöllisyysehtojen vallitessa suurimman uskottavuuden estimaattorit ovat asymptoottisesti tehokkaita, eli

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}_d(0, \boldsymbol{I}_E(\boldsymbol{\theta})^{-1}), \tag{A.3}$$

kun otoskoko $n \to \infty$, missä $\boldsymbol{\theta}$ on todellinen parametrivektorin arvo, d tämän dimensio, ja $\boldsymbol{I}_E(\boldsymbol{\theta})$ havaintojen odotettu (Fisherin) informaatiomatriisi. Merkittäviä säännöllisyysehtoja ovat vaatimus että parametrit ovat identifioituvia ($\tilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta} \Rightarrow f(x;\tilde{\boldsymbol{\theta}}) \neq f(x;\boldsymbol{\theta})$), että $\boldsymbol{\theta}$ on parametriavaruuden Θ sisäpiste, ja että $f(x;\boldsymbol{\theta})$:n määrittelyalueen (support) ei tulisi riippua parametreista $\boldsymbol{\theta}$. Säännöllisyysehdoista tarkemmin, ks. esimerkiksi [28].

Fisherin informaatio skalaariparametrille θ määritellään

$$I_E(\theta) = \mathbb{E}\left(\frac{\partial}{\partial\theta} \ln L(\theta;\boldsymbol{X})\right)^2;$$

säännöllisyysehtojen pätiessä tämä voidaan saattaa edelleen muotoon

$$I_E(\theta) = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\ln L(\theta;\boldsymbol{X})\right).$$

Vektorin $\boldsymbol{\theta}$ tapauksessa odotettu (Fisherin) informaatiomatriisi on

$$\boldsymbol{I}_E(\boldsymbol{\theta}) = -\mathbb{E}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ln L(\boldsymbol{\theta};\boldsymbol{X})\frac{\partial}{\partial\boldsymbol{\theta}^T}\ln L(\boldsymbol{\theta};\boldsymbol{X})\right) = \mathbb{E}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\ln L(\boldsymbol{\theta};\boldsymbol{X}\right),$$

millä tarkoitetaan matriisia $\boldsymbol{I}_E$ komponenteilla

$$(\boldsymbol{I}_E(\boldsymbol{\theta}))_{ij} = -\mathbb{E}\left(\frac{\partial}{\partial\theta_i}\ln L(\boldsymbol{\theta};\boldsymbol{X})\frac{\partial}{\partial\theta_j}\ln L(\boldsymbol{\theta};\boldsymbol{X})\right) = \mathbb{E}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln L(\boldsymbol{\theta};\boldsymbol{X}\right).$$

SU-estimaattorin asymptoottinen tehokkuus sisältää tuloksena sekä estimaattorin asymptoottisen normaalisuuden että estimaattorin tarkentuvuuden (missä jälkimmäinen tarkoittaa, että estimaatti $\hat{\boldsymbol{\theta}}_n$ suppenee todennäköisyyden suhteen todelliseen arvoon $\boldsymbol{\theta}$ kun $n \to \infty$). Suppenemistuloksesta (A.3) seuraa että, riittävän suurella n,

$$\hat{\boldsymbol{\theta}}_n \sim \mathrm{N}_d(\boldsymbol{\theta}, n^{-1}\boldsymbol{I}_E(\boldsymbol{\theta})^{-1}), \tag{A.4}$$

ja tätä voidaan käyttää hyväksi muodostettaessa asymptoottisia luottamusalueita $\boldsymbol{\theta}$:lle, tai, tavallisemmin, luottamusvälejä sen mille tahansa komponentille θ_i. Käytännössä on yleensä helpompaa approksimoida odotettua informaatiomatriisia $\boldsymbol{I}_E(\boldsymbol{\theta})$ havaitulla (Fisherin) informaatiomatriisilla $\boldsymbol{I}_O(\boldsymbol{\theta})$,

$$\boldsymbol{I}_O(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\ln L(\boldsymbol{\theta};x_i), \tag{A.5}$$

missä $\boldsymbol{x} = (x_1,\ldots,x_n)$ on havaittu otos eli X:n havaittu realisaatio. Havaittu informaatiomatriisi suppenee odotettuun suurten lukujen lain nojalla, ja on esitetty, että joissain tilanteissa sen käyttö johtaa jopa tarkempiin tuloksiin kuin odotetun informaatiomatriisin käyttö (ks. [1]). Molemmat versiot informaatiomatriisista riippuvat tuntemattomista parametreista $\boldsymbol{\theta}$, ja matriisien laskemiseksi on tarpeen korvata parametrivektori $\boldsymbol{\theta}$ estimaatillaan $\hat{\boldsymbol{\theta}}$.

A.2 Asymptoottiseen normaalisuuteen perustuvat luottamusvälit

Tuloksesta (A.4) seuraa, että standardoitu muuttuja

$$Z = \frac{\hat{\theta}_i - \theta_i}{\mathrm{se}(\hat{\theta}_i)} \sim \mathrm{N}(0,1), \tag{A.6}$$

missä $\mathrm{se}(\hat{\theta}_i)$ on $\hat{\theta}_i$:n asymptoottinen keskivirhe (asymptoottisen keskihajonnan estimaatti),

$$\mathrm{se}(\hat{\theta}_i) = \sqrt{\frac{1}{n}\left(\boldsymbol{I}_O(\hat{\boldsymbol{\theta}})\right)_{ii}}.$$

(A.6) on siis asymptoottinen tulos, jonka tulkitaan pätevän riittävän suurilla n. Yhtälön (A.6) perusteella voidaan muodostaa testi nollahypoteesille $H_0 : \theta_i = \theta_{i,0}$ vaihtoehtoista hypoteesia $H_1 : \theta_i \neq \theta_{i,0}$ vastaan, missä $\theta_{i,0}$ on jokin kiinnostuksen kohteena oleva parametrin arvo. Asymptoottiselle testille merkittävyydellä α nollahypoteesi hylätään, jos $Z \geq \Phi^{-1}(1-\alpha/2)$, missä Φ on standardin normaalijakauman kertymäfunktio.

Testiin perustuen voidaan rakentaa myös luottamusvälit parametrille θ_i: asymptoottinen luottamusväli luottamustasolla $1-\alpha$ muodostuu niistä parametriarvoista $\theta_{i,0}$, joilla nollahypoteesia H_0 ei hylätä, ja on siis

$$\left(\hat{\theta}_i - z_{\alpha/2}\mathrm{se}(\hat{\theta}_i), \hat{\theta}_i + z_{\alpha/2}\mathrm{se}(\hat{\theta}_i)\right),$$

missä $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ on jakauman $\mathrm{N}(0,1)$ $(1-\alpha/2)$-kvantiili.

Tässä esityksessä omaksutaan Colesin [1] pragmaattinen lähestymistapa, ja käytetään asymptoottisia tuloksia suoraan hyväksi pitämällä niitä (hyväksyttävinä) approksimaatioina, joiden tarkkuus paranee otoskoon n kasvaessa.

A.2.1 Delta-menetelmä

Usein on tarvetta tarkastella parametrien $\boldsymbol{\theta}$ SU-estimaattorin $\hat{\boldsymbol{\theta}}$ lisäksi jotakin $\boldsymbol{\theta}$:n funktiota. Olkoon $\phi = g(\boldsymbol{\theta})$ tällainen funktio. Tällöin funktion ϕ suurimman uskottavuuden estimaattori saadaan yksinkertaisesti sijoittamalla parametrivektorin paikalle sen SU-estimaattori $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, eli $\hat{\phi} = g(\hat{\boldsymbol{\theta}})$.

Olkoon edelleen d-ulotteisen parametrivektorin $\boldsymbol{\theta}$ SU-estimaatin $\hat{\boldsymbol{\theta}}$ asymptoottinen kovarianssimatriisi $\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{I}_E(\boldsymbol{\theta})^{-1}$. Tällöin funktion $\phi = g(\boldsymbol{\theta})$ suurimman uskottavuuden estimaattorille pätee

$$\hat{\phi} \dot{\sim} N(\phi, \boldsymbol{V}_\phi),$$

missä

$$\boldsymbol{V}_\phi = \nabla g(\boldsymbol{\theta})^T \boldsymbol{V}_{\boldsymbol{\theta}} \nabla g(\boldsymbol{\theta})$$

ja gradientti ∇g on

$$\nabla g(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial\theta_1} g(\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial\theta_d} g(\boldsymbol{\theta})\right),$$

evaluoituna pisteessä $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Tätä kutsutaan *delta-menetelmäksi*. Menetelmä mahdollistaa luottamusvälien muodostamisen estimaattorin $\hat{\phi}$ asymptoottiseen normaalisuuteen perustuen, samalla tavalla kuin edellisessä osiossa.

Liite B

Uskottavuusosamäärätesti

Tarkastellaan nollahypoteesin $H_0 : \boldsymbol{\theta} \in \Theta_0$ testaamista vaihtoehtoista hypoteesiä $H_0 : \boldsymbol{\theta} \in \Theta_0^c$ vastaan, missä $\Theta_0 \subset \Theta$. Uskottavuusosamäärätestin testisuure määritellään

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}; \boldsymbol{x})}{\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \boldsymbol{x})},$$

missä $\boldsymbol{x} = (x_1, \ldots, x_n)$ on iid satunnaismuuttujien $\boldsymbol{X}$ realisaatio kuten aiemmin, ja soveltuvien säännöllisyysehtojen oletetaan edelleen pätevän. Voidaan osoittaa, että nollahypoteesin H_0 vallitessa

$$-2 \ln \lambda(\boldsymbol{x}) \sim \chi^2_\nu,$$

missä χ^2-jakauman vapausasteiden määrä ν on olennaisesti ehdon $\boldsymbol{\theta} \in \Theta$ määrittämien vapaiden parametrien lukumäärä vähennettynä nollahypoteesin $\boldsymbol{\theta} \in \Theta_0$ määrittämällä vapaiden parametrien määrällä.

Oletetaan, että d-dimensioinen parametrivektori $\boldsymbol{\theta}$ voidaan osittaa kahteen komponenttiin, $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$, missä $\boldsymbol{\theta}^{(1)}$ on kiinnostuksen kohteena oleva k-dimensioinen vektori, ja $\boldsymbol{\theta}^{(2)}$ vastaa jäljelle jääviä $d - k$ komponenttia. Pyrkimyksenä on testata nollahypoteesia $H_0 : \boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}_0^{(1)}$ vaihtoehtoista hypoteesia $H_0 : \boldsymbol{\theta}^{(1)} \neq \boldsymbol{\theta}_0^{(1)}$ vastaan. Merkitään uskottavuusfunktiota $L(\boldsymbol{\theta}; \boldsymbol{X}) = L(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}; \boldsymbol{X})$. Uskottavuusosamäärätestin testisuureelle pätee nyt

$$-2 \ln \lambda(\boldsymbol{x}) = -2 \left(\ln L(\boldsymbol{\theta}_0^{(1)}, \hat{\boldsymbol{\theta}}_0^{(2)}; \boldsymbol{x}) - \ln L(\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}; \boldsymbol{x}) \right) \sim \chi^2_k,$$

missä $\hat{\boldsymbol{\theta}}^{(1)}$ ja $\hat{\boldsymbol{\theta}}^{(2)}$ ovat komponenttivektorien $\boldsymbol{\theta}^{(1)}$ ja $\boldsymbol{\theta}^{(2)}$ rajoittamattomat SU-estimaatit[1] ja $\hat{\boldsymbol{\theta}}_0^{(2)}$ on komponenttivektorin $\boldsymbol{\theta}^{(2)}$ rajoitettu SU-estimaatti nollahypoteesin H_0 pätiessä. Yllä oleva yhtälö voidaan kirjoittaa myös muodossa

$$D(\boldsymbol{\theta}_0^{(1)}) = 2 \left(l(\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}; \boldsymbol{x}) - l(\boldsymbol{\theta}_0^{(1)}, \hat{\boldsymbol{\theta}}_0^{(2)}; \boldsymbol{x}) \right),$$

missä suuretta $D = -2 \ln \lambda$ kutsutaan yleensä devianssifunktioksi tai devianssiksi. Nollahypoteesi H_0 hylätään, mikäli $D > c_{k,\alpha}$, missä $c_{k,\alpha}$ on χ^2_k-jakauman $(1 - \alpha)$-kvantiili.

[1] $(\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}) = \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \boldsymbol{X}) = \arg\max_{(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) \in \Theta} L((\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}); \boldsymbol{X})$

B.1 Luottamusvälit ja profiiliuskottavuus

Luottamusjoukko parametrivektorille $\boldsymbol{\theta}$ ja luottamusvälit sen mille tahansa komponentille voidaan perustaa uskottavuusosamäärätestiin. Asymptoottinen luottamus*joukko* vektorille $\boldsymbol{\theta}^{(1)}$ luottamustasolla $1-\alpha$ muodostuu niistä arvoista $\boldsymbol{\theta}_0^{(1)}$, joilla nollahypoteesia $H_0 : \boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}_0^{(1)}$ ei hylätä, eli on

$$\left\{\boldsymbol{\theta}_0^{(1)} : D(\boldsymbol{\theta}_0^{(1)}) \leq c_{k,\alpha}\right\},$$

tai

$$\left\{\boldsymbol{\theta}_0^{(1)} : l\left(\boldsymbol{\theta}_0^{(1)}, \hat{\boldsymbol{\theta}}_0^{(2)}; \boldsymbol{x}\right) \geq l\left(\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}; \boldsymbol{x}\right) - \frac{1}{2}c_{k,\alpha}\right\}.$$

Erityisesti, jos $k = 1$ – kuten silloin kun ollaan kiinnostuneita vain yhdestä parametrivektorin komponentista $\theta^{(1)}$ kerrallaan – saadaan komponentin $\theta^{(1)}$ luottamus*väliksi*

$$\left\{\theta_0^{(1)} : D(\theta_0^{(1)}) \leq c_{k,\alpha}\right\},$$

eli

$$\left\{\theta_0^{(1)} : l\left(\theta_0^{(1)}, \hat{\boldsymbol{\theta}}_0^{(2)}; \boldsymbol{x}\right) \geq l\left(\hat{\theta}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}; \boldsymbol{x}\right) - \frac{1}{2}c_{1,\alpha}\right\}.$$

Käyrää $\left(\theta_0^{(1)}, l(\theta_0^{(1)}, \hat{\boldsymbol{\theta}}_0^{(2)}; \boldsymbol{x})\right)$ kutsutaan parametrin $\theta_0^{(1)}$ *profiili(log-)uskottavuuskäyräksi* tai *profiili(log-)uskottavuudeksi*. Merkitään profiiliuskottavuutta komponentille θ_i lyhyesti

$$l_p(\theta_i) = \max_{\boldsymbol{\theta}_{-i}} l\left(\theta_i, \boldsymbol{\theta}_{-i}; \boldsymbol{x}\right) = l\left(\theta_i, \hat{\boldsymbol{\theta}}_{-i}; \boldsymbol{x}\right),$$

missä $\boldsymbol{\theta}_{-i}$ tarkoittaa parametrivektorin kaikkia muita paitsi i. komponenttia, ja $\hat{\boldsymbol{\theta}}_{-i} = \hat{\boldsymbol{\theta}}_{-i}(\theta_i)$ tarkoittaa vektorin $\boldsymbol{\theta}_{-i}$ suurimman uskottavuuden estimaattia kun θ_i on kiinnitetty. Profiiliuskottavuus parametrille θ_i on siis kullakin θ_i:n arvolla kaikkien muiden parametrivektorin komponenttien suhteen maksimoitu log-uskottavuus.

Oletetaan seuraavaksi, että d-dimensioinen parametrivektori $\boldsymbol{\theta}$ voidaan osittaa kahteen komponenttiin, $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$ kuten aiemmin, missä $\boldsymbol{\theta}^{(1)}$ on k-dimensioinen vektori ja $\boldsymbol{\theta}^{(2)}$ vastaa jäljelle jääviä $d-k$ komponenttia. Profiiliuskottavuus vektorille $\boldsymbol{\theta}^{(1)}$ voidaan nyt kirjoittaa lyhyesti

$$l_p(\boldsymbol{\theta}^{(1)}) = \max_{\boldsymbol{\theta}^{(2)}} l\left(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}; \boldsymbol{x}\right) = l\left(\boldsymbol{\theta}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}; \boldsymbol{x}\right),$$

missä merkinnöillä on sama tulkinta kuin skalaaritapauksessa.

B.2 Mallin valinta

Sisäkkäisten mallien tapauksessa – eli silloin kun toinen malli muodostaa yleisemmän mallin erikoistapauksen tiettyjen parametriarvojen ollessa rajoitettuja – mallin valinta voidaan perustaa uskottavuusosamäärätestiin. Tarkastellaan

kahta mallia, $\mathcal{M}_0$ ja $\mathcal{M}_1$: Olkoon mallia $\mathcal{M}_1$ vastaava (d-ulotteinen) parametrivektori $\boldsymbol{\theta}$, ja olkoon $\mathcal{M}_0$ mallin $\mathcal{M}_1$ alimalli, joka saadaan rajoittamalla $\boldsymbol{\theta}$:n k ensimmäistä komponenttia nolliksi. Parametrivektori voidaan siis kirjoittaa $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$, missä ensimmäinen komponentti vastaa nollavektoria, $\boldsymbol{\theta}^{(1)} = \mathbf{0}$, mallissa $\mathcal{M}_0 \subset \mathcal{M}_1$.

Olkoot malleja vastaavat maksimoidut log-uskottavuudet edelleen $l_1(\mathcal{M}_1)$ ja $l_0(\mathcal{M}_0)$. Testisuure on nyt

$$D = 2(l_1(\mathcal{M}_1) - l_0(\mathcal{M}_0)) \sim \chi^2_k,$$

ja yksinkertaisempi malli $\mathcal{M}_0$ hylätään mallin $\mathcal{M}_1$ eduksi merkittävyystasolla α, jos $D > c_{k,\alpha}$, missä $c_{k,\alpha}$ on χ^2_k-jakauman $(1-\alpha)$-kvantiili.

B.2.1 Informaatiokriteerit

Kuten edellä mainittiin, uskottavuusosamäärätesti soveltuu sisäkkäisten mallien vertailuun. Kun kuitenkin yleisemmin halutaan verrata malleja, jotka eivät ole sisäkkäisiä ja joissa saattaa olla hyvinkin eri määrä parametreja, formaalia tilastollista testiä mallinvalinnan perusteeksi ei yleisessä tapauksessa ole olemassa.

Ei-sisäkkäisten mallien vertailuun käytetään yleensä ns. *informaatiokriteerejä*, joiden perusajatuksena on mitata tarkasteltavan mallin sopivuutta sen saavuttamaan (log-)uskottavuuteen perustuen, mutta asettaa mallille sakko, joka rankaisee kompleksisuudesta ja on yleensä verrannollinen mallin sisältämien parametrien määrään. Sakon tavoitteena on tehdä eri (parametrimäärän sisältävistä) malleista keskenään mahdollisimman vertailukelpoisia. Eräs tunnetuimmista informaatiokriteereistä on Akaiken informaatiokriteeri (AIC). [43] Oletetaan, että tarkasteltavana on m mallia, $\mathcal{M}_1, \ldots, \mathcal{M}_m$, ja että mallilla j on k_j parametria, merkitään $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jk_j})^T$, ja sitä vastaava log-uskottavuus on $l_j(\boldsymbol{\theta}_j; \boldsymbol{x})$. AIC mallille j määritellään nyt

$$\mathrm{AIC}(\mathcal{M}_j) = 2l_j(\hat{\boldsymbol{\theta}}_j; \boldsymbol{x}) - 2k_j$$

missä $\hat{\boldsymbol{\theta}}$ on parametrivektorin SU-estimaatti. Malleista valitaan lähestymistavassa se, joka minimoi luvun AIC eli sakotetun log-uskottavuuden. Sakkona toimii Akaiken informaatiokriteerissä suoraan malliparametrien lukumäärä. Toinen yleisesti käytetty informaatiokriteeri on (Schwartzin) Bayesilainen informaatiokriteeri BIC [44], jossa sakkona on $k_j \ln(n)$, missä n on otoskoko eli havaintojen lukumäärä. Informaatiokriteerit eroavat yleisesti toisistaan sakkotermin muodon suhteen.

Liite C

Poisson-prosesseista

Esitetään lyhyesti tavallisen Poisson-prosessin ja epähomogeenisen Poisson-prosessin määritelmät $\mathbb{R}_+ = [0, \infty)$:ssä. Esitys perustuu lähteeseen [15]; ks. myös [5]. Aloitetaan määrittelemällä lukumäärämuuttuja ja laskuriprosessi.

Olkoon annettu todennäköisyyskenttä $(\Omega, \mathscr{F}, \mathbb{P})$. Satunnaistapahtumien (esimerkiksi vahinkojen) lukumäärää kiinteällä aikavälillä voidaan kuvata ei-negatiivisella satunnaismuuttujalla: Satunnaismuuttujaa $N : (\Omega, \mathscr{F}) \to (\mathbb{R}_+, \mathscr{B}_+)$, missä $\mathscr{B}_+$ tarkoittaa $\mathbb{R}_+$:n Borel-sigma-algebraa, kutsutaan *lukumäärämuuttujaksi*, jos $\mathbb{P}(N \in \{0, 1, 2, \ldots\}) = 1$. Merkitään tapahtumien lukumäärän aikavälillä $(0, t]$ laskevaa lukumäärämuuttujaa $N(t)$. Olkoon vastaavasti $N(t, u)$ aikavälillä $(t, u]$, $0 \leq t < u$, sattuneiden tapahtumien lukumäärä, jolloin $N(t, u) = N(u) - N(t)$.

Kun tarkastellaan satunnaisilmiön kehitystä ajan suhteen (esimerkiksi vahinkojen sattumista ajassa), tarvitaan sopivaa stokastista prosessia.

Määritelmä C.1 (Laskuriprosessi)
Stokastinen prosessi $N = (N(t))_{t \geq 0}$ on laskuriprosessi*, jos $N(t)$ on satunnaismuuttuja ja seuraavat ehdot pätevät kaikilla $t \geq 0$:*

(i) $\mathbb{P}(N(0) = 0) = 1$,

(ii) $N(t)$ *on lukumäärämuuttuja,*

(iii) prosessin polut ovat oikealta jatkuvia ja niillä on vasemmanpuoleiset rajaarvot (cádlág); toisin sanoen kuvaus $f_\omega : [0, \infty) \to \mathbb{R}$, $f_\omega(t) = N(t)(\omega)$ on cádlág $\forall\, \omega \in \Omega$,

(iv) $\mathbb{P}(N(t) - N(t-) = 0 \ \vee \ 1, \forall\, t > 0) = 1$, *missä* $N(t-) = \lim_{h \to 0} N(t - h)$.

Laskuriprosessi on siis aina kokonaislukuarvoinen (kohta (ii)) ja prosessin hypyt ovat aina ykkösen suuruisia (kohta (iv) – ts. prosessi on yksinkertainen).

Poisson-jakautunut satunnaismuuttuja on lukumäärämuuttujan erikoistapaus, ja Poisson-prosessi on laskuriprosessin erikoistapaus. Satunnaismuuttujan K sanotaan olevan *Poisson-jakautunut* parametrilla $\lambda > 0$, $K \sim \text{Poi}(\lambda)$, jos

$$\mathbb{P}(K = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots.$$

Mikäli $\lambda = 0$, asetetaan $\mathbb{P}(K = 0) = 1$.

Määritelmä C.2 (Poisson-prosessi)
Stokastinen prosessi $N = (N(t))_{t\geq 0}$ *on* Poisson-prosessi *intensiteetillä* $\lambda \geq 0$, *jos seuraavat ehdot täyttyvät:*

(i) $(N(t))_{t\geq 0}$ *on laskuriprosessi,*

(ii) $N(u-t) \sim \mathrm{Poi}(\lambda(u-t))$ *kaikilla* $0 \leq t < u$,

(iii) mielivaltaisille ajanhetkille $0 \leq t_1 < u_1 \leq t_2 < u_2 \leq \ldots \leq t_n < u_n$, $n \in \mathbb{N}$, *prosessin lisäykset* $N(t_1, u_1), \ldots, N(t_n, u_n)$ *ovat riippumattomia.*

Poisson-prosessi (tai homogeeninen Poisson-prosessi) on siis stokastinen prosessi, joka on laskuriprosessi riippumattomilla ja stationaarisillä lisäyksillä, ja jolle pätee $N(t) \sim \mathrm{Poi}(\lambda t)$ jokaisella $t > 0$, missä $N(t)$ on $\mathscr{F}$-mitallinen satunnaismuuttuja $\Omega \to \{0, 1, 2, \ldots\}$.

Vaihtoehtoinen tapa määritellä Poisson-prosessi (ja laskuriprosessi) on seuraava: Olkoon $(T_n)_{n\geq 0}$ jono ei-negatiivisia satunnaismuuttujia, joille pätee $0 \leq T_1 \leq T_2 \leq \cdots$ (melkein varmasti). Tällöin ehdosta

$$N(t) = \sup\{n \geq 1 \mid T_n \leq t\}, \quad t \geq 0,$$

määritelty stokastinen prosessi $(N(t))_{t\geq 0}$ on laskuriprosessi $\mathbb{R}_+$:ssa (yllä asetetaan tyhjän joukon supremum nollaksi, $\sup A = 0$ jos $A = \emptyset$). Satunnaismuuttujat T_n voidaan tulkita puhtaan hyppyprosessin $(N(t))$ hyppyjen sattumisajoiksi. Jos muuttujille T_n pätee $T_n = \xi_1 + \cdots + \xi_n$, $n \geq 1$, missä $\xi, \xi_1, \xi_2, \ldots$ ovat iid eksponenttijakautuneita satunnaismuuttujia parametrilla λ,

$$\mathbb{P}(\xi \leq x) = 1 - e^{-\lambda x}, \quad x \geq 0,$$

niin $(N(t))$ on Poisson-prosessi. Satunnaismuuttujat $\xi_n = T_n - T_{n-1}$ (missä $\xi_1 = T_1 - 0 = T_1$) ilmaisevat hyppyjen välisen ajan, ja niitä kutsutaan tavanomaisesti odotusajoiksi tai saapumisten välisiksi ajoiksi (inter-arrival times). Yleisemmin, laskuriprosessia, jonka generoi mikä tahansa yo. muotoa oleva iid satunnaismuuttujien (ξ_i) (ei välttämättä eksponenttijakautuneiden) muodostama summaprosessi (T_n), kutsutaan uusiutumisprosessiksi (renewal process).

Kun homogeenisen Poisson-prosessin vakiointensiteetti λ korvataan deterministisellä funktiolla $\lambda(\cdot)$, saadaan epähomogeeninen Poisson-prosessi. Olkoon intensiteettimitta tai kumulatiivinen intensiteetti $\Lambda : [0, \infty) \to [0, \infty)$, $\Lambda(t) = \int_0^t \lambda(s)\, \mathrm{d}s$, missä intensiteetti $\lambda(\cdot)$ on ei-negatiivinen funktio. Koska λ on ei-negatiivinen, on Λ kasvava.

Määritelmä C.3 (Epähomogeeninen Poisson-prosessi)
Stokastinen prosessi $N = (N(t))_{t\geq 0}$ *on* epähomogeeninen Poisson-prosessi *intensiteettimitalla* Λ, *jos seuraavat ehdot täyttyvät:*

(i) $(N(t))_{t\geq 0}$ *on laskuriprosessi,*

(ii) $N(u-t) \sim \mathrm{Poi}(\Lambda(u) - \Lambda(t))$ *kaikilla* $0 \leq t < u$,

(iii) mielivaltaisille ajanhetkille $0 \leq t_1 < u_1 \leq t_2 < u_2 \leq \ldots \leq t_n < u_n$, $n \in \mathbb{N}$, *prosessin lisäykset* $N(t_1, u_1), \ldots, N(t_n, u_n)$ *ovat riippumattomia.*

Intensiteettimitalle pätee siis $\Lambda(u) - \Lambda(t) = \int_t^u \lambda(s)\,\mathrm{d}s$. Kun valitaan $\Lambda(t) = \lambda t$, kaikilla $t \geq 0$, saadaan homogeeninen Poisson-prosessi.

Oletetaan seuraavaksi, että λ on positiivinen – kuten yleensä käytännössä on – jolloin $\Lambda(t)$ on aidosti kasvava, ja pätee $\Lambda(\Lambda^{-1}(t)) = \Lambda^{-1}(\Lambda(t)) = t$. Tällöin epähomogeeninen Poisson-prosessi N intensiteettimitalla Λ voidaan aina muuntaa homogeeniseksi Poisson-prosessiksi intensiteetillä 1:

Propositio C.4 (Prosessin aikamuunnos)
Olkoon N epähomogeeninen Poisson-prosessi intensiteettimitalla Λ, missä Λ on aidosti kasvava. Määritellään $\tilde{N} = N(\Lambda^{-1}(t))$, $t \geq 0$; tällöin $\tilde{N}$ on homogeeninen Poisson-prosessi intensiteetillä 1.

Todistus. Kiinteällä $t > 0$ ja $k \geq 0$

$$\mathbb{P}(\tilde{N}(t) = n) = \mathbb{P}(N(\Lambda^{-1}(t)) = n) = e^{-\Lambda(\Lambda^{-1}(t))}\frac{(\Lambda(\Lambda^{-1}(t)))^n}{n!} = e^{-t}\frac{t^n}{n!},$$

eli $\tilde{N}(t) \sim \mathrm{Poi}(t)$ (ts. intensiteetti on λt, missä $\lambda = 1$). Täytyy näyttää vielä lisäyksien riippumattomuus ja stationaarisuus. Nyt prosessin $\tilde{N}$ lisäykset ovat riippumattomia suoraan määritelmän mukaan, ja, kun $0 < s < t$,

$$\begin{aligned}\mathbb{P}(\tilde{N}(t) - \tilde{N}(s) = n) &= \mathbb{P}(N(\Lambda^{-1}(t)) - N(\Lambda^{-1}(s)) = n)\\ &= e^{-\left\{\Lambda(\Lambda^{-1}(t)) - \Lambda(\Lambda^{-1}(s))\right\}}\frac{\left\{\Lambda(\Lambda^{-1}(t)) - \Lambda(\Lambda^{-1}(s))\right\}^n}{n!}\\ &= e^{-(t-s)}\frac{(t-s)^n}{n!},\end{aligned}$$

mistä stationaarisuus seuraa. □

Liite D

Pisteprosesseista

Täsmennetään tässä liitteessä osion 1.6 pisteprosessien käsittelyä joiltakin osin. Seuraavan esityksen ei ole tarkoitus olla yhtenäinen esitys aiheesta, vaan lyhyt kokoelma täydentäviä tuloksia sekä muutamia (havainnollistavia) todistuksia. Aloitetaan asetelman käsittely satunnaismittojen kontekstissa kokonaisuuden hahmottamiseksi, vaikka tätä yleisyystasoa ei jatkossa suoranaisesti tarvitakaan. Esitys noudattaa lähteitä [3, luku 3] ja [17, luku 1 ja liite A].

Olkoon E lokaalisti kompakti Hausdorff-avaruus jonka topologialla on numeroituva kanta.[1] Avaruus E on varustettu E:n osajoukkojen muodostamalla Borel-σ-algebralla $\mathscr{E}$, ts. avoimien joukkojen generoimalla σ-algebralla. Olkoon edelleen $\mathscr{B}$ rajoitettujen (suhteellisesti kompaktien) Borel-joukkojen muodostama perhe $\mathscr{E}$:ssä. Merkitään $C_K(E)$:llä jatkuvien, reaaliarvoisten E:n funktioiden joukkoa kompaktilla tukijoukolla.[2]

Määritelmä D.1 (Radon-mitta)
Mitta μ avaruudella E on Radon-mitta*, jos $\mu(K) < \infty$ jokaisella kompaktilla joukolla $K \in \mathscr{E}$.*

Pätee: μ on Radon-mitta jos ja vain jos $|\mu(f)| = |\int f d\mu| < \infty$ kaikilla $f \in C_K(E)$.

Muistetaan osiosta 1.6 seuraavat määritelmät. Kun $x \in E$, asetetaan

$$\delta_x(A) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A, \end{cases} \qquad \text{kun } A \in \mathscr{E}.$$

Mittaa δ_x kutsutaan Dirac-mitaksi (Diracin delta-mitaksi) tai pistemassaksi. Olkoon $\{x_i : i \geq 1\}$ numeroituva joukko pisteitä tila-avaruudessa E. Laskurimitta E:ssä on mitta m, joka on seuraavaa muotoa:

$$m(A) = \sum_{i=1}^{\infty} \delta_{x_i}(A) = \text{card}\{i : x_i \in A\}, \quad A \in \mathscr{E}.$$

[1] Siis E on Hausdorff, jokaisella pisteellä $x \in E$ on olemassa kompakti ympäristö, ja on olemassa avoimet joukot $\{G_i, i \geq 1\}$ s.e. mikä tahansa avaruuden avoin joukko G voidaan kirjoittaa yhdisteenä $G = \bigcup_{\alpha \in I} G_\alpha$, missä I on äärellinen tai numeroituva indeksijoukko.

[2] Siis $f \in C_K(E)$ tarkoittaa, että on olemassa kompakti joukko K ja jatkuvalle funktiolle f pätee $f(x) = 0$ kun $x \in K^c$.

Siis laskurimitan määrittelevä ominaisuus on, että $m(A) \in \bar{\mathbb{N}} \cup \{0\}$ kaikilla A. Jos $m(K) < \infty$ jokaiselle kompaktille joukolle $K \in \mathscr{E}$ eli m on Radon, nimitetään mittaa m edelleen pistemitaksi.

Olkoon $\mathbb{M}(E)$ Radon-mittojen joukko E:llä, ja $M_p(E) \subset \mathbb{M}(E)$ kaikkien tila-avaruuden E pistemittojen muodostama osajoukko. $\mathbb{M}(E)$:llä määritellään σ-algebra

$$\mathscr{M}(E) = \sigma\left(m \to m(f) : f \in C_K(E)\right);$$

$\mathscr{M}(E)$ generoituu koordinaattikuvausten $m \to m(f) = \int_E f dm$ kautta, kun f käy läpi jatkuvien, kompaktin tukijoukon omaavien funktioiden joukon $C_K(E)$. $\mathscr{M}(E)$ on siis pienin $\mathbb{M}(E)$:n osajoukkojen muodostama σ-algebra, joka tekee kuvauksista $m \to m(f)$ $\mathbb{M}(E) \to \mathbb{R}$ mitallisia kaikilla $f \in C_K(E)$. Vaihtoehtoisesti $\mathscr{M}$ voidaan karakterisoida joukkojen kautta:

$$\mathscr{M}(E) = \sigma\left\{\{m \in \mathbb{M}(E) : m(f) \in B\}, f \in C_K(E), B \in \mathscr{B}\right\}.$$

Pisteprosessien avaruus $M_p(E)$ kuuluu $\mathscr{M}(E)$:hen, ja voidaan ottaa $\mathscr{M}_p(E) = \mathscr{M}(E) \cap M_p(E) := \{A \cap M_p(E) : A \in \mathscr{M}(E)\}$.

Määritelmä D.2 (Satunnaismitta ja pisteprosessi)
Olkoon $(\Omega, \mathscr{F}, \mathbb{P})$ *todennäköisyysavaruus.*

(a) Satunnaismitta *tila-avaruudella* E *on mitallinen kuvaus*

$$M : (\Omega, \mathscr{F}) \to (\mathbb{M}(E), \mathscr{M}(E)).$$

(b) Pisteprosessi *tila-avaruudella* E *on mitallinen kuvaus*

$$N : (\Omega, \mathscr{F}) \to (M_p(E), \mathscr{M}_p(E)).$$

Satunnaismitta on määritelmän mukaan avaruuden $\mathbb{M}(E)$ satunnainen elementti, ja pisteprosessi vastaavasti avaruuden $M_p(E)$ satunnainen elementti. Mikäli satunnaismitta ottaa kaikki arvonsa joukossa $\{0, 1, \ldots, \infty\}$, yhtyvät yllä olevat määritelmät. Näin ollen tässä esityksessä tullaan käyttämään "pisteprosessia" ja "satunnaismittaa" osittain synonyymeinä, etenkin Poisson-satunnaismitan kohdalla.

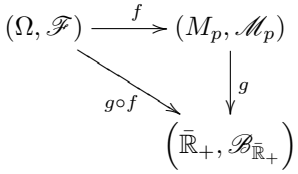
Tarkastellaan pisteprosessien avaruutta $M_p(E)$ ja vastaavaa σ-algebraa hieman eksplisiittisemmin. $M_p(E)$ on kaikkien tila-avaruudessa E määriteltyjen pistemittojen joukko, $M_p(E) = \{m \in M : m(A) \in \mathbb{N}_0 \forall A \in \mathscr{B}\}$. $M_p(E)$:n osajoukkojen generoima sigma-algebra $\mathscr{M}_p(E)$ voidaan [3]:n mukaisesti ottaa pienimmäksi sigma-algebraksi, joka sisältää kaikki muotoa $\{m \in M_p(E) : m(A) \in B\}$, $A \in \mathscr{E}$, $B \in \mathscr{B}([0, \infty])$ olevat joukot. Toisin sanottuna, $\mathscr{M}_p(E)$ on pienin sigma-algebra, joka tekee kaikki evaluaatiokuvaukset $m \to m(A) : (M_p(E) \to [0, \infty])$ mitallisiksi kaikilla $A \in \mathscr{E}$.

Kuten edellä mainittu, pisteprosessi on avaruuden $M_p(E)$ satunnainen elementti; ts. satunnainen elementti tai funktio, joka ottaa arvoikseen pistemittoja. Suoraan yllä olevan määritelmän perusteella voi olla vaikea osoittaa, että tietty kuvaus $N : \Omega \to M_p(E)$ on pisteprosessi. Seuraava tulos ([3, Proposition 3.1]) antaa mukavamman kriteerin; proposition todistus havainnollistaa myös edellä esiintyneiden avaruuksien välisiä suhteita.

Propositio D.3

$$N \text{ on pisteprosessi} \iff \text{kuvaus } \omega \to N(\omega, A) \text{ on mitallinen } (\Omega, \mathscr{F}) \to \left([0,\infty], \mathscr{B}([0,\infty])\right) \forall\, A \in \mathscr{E}.$$

Todistus. "⇒": Jos N on pisteprosessi, niin kuvaus $f : \omega \to N(\omega, \cdot)$ on mitallinen avaruudesta $(\Omega, \mathscr{F}) \to (M_p(E), \mathscr{M}_p(E))$. Lisäksi evaluaatiokuvaus $g_A : m \to m(A)$ on mitallinen $(M_p(E), \mathscr{M}_p(E)) \to ([0,\infty], \mathscr{B}([0,\infty]))$ σ-algebran $\mathscr{M}_p(E)$ määritelmän perusteella. Täten kuvaus $\omega \to N(\omega, A)$ on mitallinen $(\Omega, \mathscr{F}) \to ([0,\infty], \mathscr{B}([0,\infty]))$ mitallisten kuvausten f ja g yhdisteenä $g \circ f$.



"⇐": Oletetaan, että kuvaus $\omega \to N(\omega, A)$ on mitallinen, ts. $\{\omega \in \Omega : N(\omega, A) \in B\} \in \mathscr{F}$ kaikilla $B \in \mathscr{B}([0,\infty])$ ja $A \in \mathscr{E}$. Määritellään joukko

$$\mathscr{G} = \{G \in \mathscr{M}_p(E) : N^{-1}(G) \in \mathscr{F}\},$$

missä siis $N^{-1}(G) = \{\omega : N(\omega, \cdot) \in G\}$. $\mathscr{G}$ on σ-algebra, ja sisältää kaikki muotoa $\{m : m(A) \in B\}$ olevat joukot, sillä

$$\begin{aligned} N^{-1}(\{m : m(A) \in B\}) &= \{\omega : N(\omega, \cdot) \in \{m : m(A) \in B\}\} \\ &= \{\omega : N(\omega, A) \in B\} \in \mathscr{F} \end{aligned}$$

alun oletuksen mukaan. Täten saadaan

$$\mathscr{G} \supset \sigma\left(\{\{m : m(A) \in B\}, A \in \mathscr{E}, B \in \mathscr{B}([0,\infty])\}\right) = \mathscr{M}_p(E).$$

Siis kuvaus $\omega \to N(\omega, A)$ on mitallinen $(\Omega, \mathscr{F}) \to (M_p(E), \mathscr{M}_p(E))$, ja N on pisteprosessi. □

Tulos sanoo, että N on pisteprosessi jos ja vain jos $N(A)$ on (laajennettu, reaaliarvoinen) satunnaismuuttuja[3] jokaiselle $A \in \mathscr{E}$ — pisteprosessia voidaan siis ajatella kokoelmana $\{N(A) : A \in \mathscr{E}\}$ laajennettuja satunnaismuuttujia $N(A)$. Itse asiassa sen osoittamiseksi, että N on pisteprosessi, ei tarvitse näyttää että $\omega \to N(\omega, A)$ on mitallinen kaikilla A, vaan riittää osoittaa mitallisuus joukoille A sopivassa rajatussa luokassa, esimerkiksi rajoitetuille suorakulmioille kun tila-avaruus E on euklidinen. Tämä on seuraavan proposition sisältö.

Propositio D.4
Olkoon $\mathscr{T}$ avaruuden $\mathscr{E}$ kompakteja osajoukkoja, joille pätee

(i) *$\mathscr{T}$ on* Π*-systeemi.*[4]

[3] Satunnaismuuttuja X on laajennettu (extended), kun $\mathbb{P}(X = \infty) > 0$, eli satunnaismuuttuja saa arvon ∞ positiivisella todennäköisyydellä.

[4] $\mathscr{T} \subset \mathscr{E}$ on Π-systeemi, jos $\mathscr{T}$ on suljettu äärellisten leikkausten suhteen: ts. jos $A, B \in \mathscr{T}$, niin $A \cap B \in \mathscr{T}$.

(ii) $\sigma(\mathscr{T}) = \mathscr{E}$.

(iii) *Joko* (a) *on olemassa* $E_n \in \mathscr{T}$, $E_n \uparrow E$, *tai* (b) *on olemassa* E*:n ositus* $\{E_n\}$, *jolle* $\sum_j E_j = E$ *ja* $E_n \in \mathscr{T}$.

Tällöin N *on pisteprosessi kentällä* $(\Omega, \mathscr{F})$ *avaruudessa* $(E, \mathscr{E})$ jos ja vain jos $\omega \to N(\omega, I)$ *on mitallinen* $\Omega \to [0, \infty)$ *jokaisella* $I \in \mathscr{T}$.

Todistus. Resnick, [3, Proposition 3.2]. □

Pisteprosessin N realisaatiot ovat pistemittoja. Täten pisteprosessin N jakauma, merkitään P_N, on mitta $\mathbb{P} \circ N^{-1} = \mathbb{P}(N \in \cdot)$, joka on määritelty pistemittojen muodostamille osajoukoille eli tapahtuma-avaruudessa $\mathscr{M}_p(E)$. Kun $A \in \mathscr{E}$, $N(A)$ on satunnaismuuttuja. Vastaavasti kun $A_1, \ldots, A_k \in \mathscr{E}$, $(N(A_i))_{i \le k}$ on satunnaisvektori. Voidaan osoittaa, että tällaisten satunnaisvektorien äärellisulotteisten jakaumien joukko määrää yksikäsitteisesti pisteprosessin todennäköisyyslain $P_N = \mathbb{P} \circ N^{-1}$:

Propositio D.5
Olkoon N *pisteprosessi avaruudessa* $(E, \mathscr{E})$ *ja oletetaan että* $\mathscr{T}$ *täyttää proposition D.4 ehdot. Määritellään massafunktiot*

$$P_{I_1,\ldots,I_k}(n_1, \ldots, n_k) = \mathbb{P}\left(N(I_1) = n_1, \ldots, N(I_k) = n_k\right)$$

kun $I_i \in \mathscr{T}$, $n_i \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$, $1 \le i \le k$. *Tällöin* P_N *määräytyy yksikäsitteisesti kun tunnetaan*

$$\{P_{I_1,\ldots,I_k},\ k = 1, 2, \ldots : I_j \in \mathscr{T}\ \forall\, 1 \le j \le k\}.$$

Todistus. Ks. [3, Proposition 3.4] ja [17, kappale 1.2]. □

Tulos ei kuitenkaan vielä sano paljon siitä, kuinka *rakentaa* pisteprosesseja. Tavanomainen tapa pisteprosessin määrittelemiseksi on seuraava: Olkoon $\{X_n : n \ge 1\}$ avaruuden E satunnaiselementtejä (jono satunnaisvektoreita) määriteltynä kentällä $(\Omega, \mathscr{F})$. Asetetaan

$$N = \sum_{i=1}^{\infty} \delta_{X_i}.$$

Tällöin

$$N(\omega, A) = \sum_{i=1}^{\infty} \delta_{X_i(\omega)}(A), \quad A \in \mathscr{E},$$

määrittelee pistemitan $\mathscr{E}$:ssä jokaisella $\omega \in \Omega$.

Seuraavaksi siirrytään käsittelemään esityksen kannalta keskeistä Poisson-pisteprosessia ja esitetetään Poisson-prosessin konstruktio. Sitä ennen otetaan käyttöön vielä seuraava käsite: Pisteprosessin *intensiteettimitta* tai *keskiarvomitta* (mean measure, first moment measure) on mitta μ, joka määritellään

$$\mu(A) = \mathbb{E}N(A) = \int_{\Omega} N(\omega, A)\mathbb{P}(d\omega),$$

kun $A \in \mathscr{E}$. N on integroituva, jos $\mu \in \mathbb{M}$. Olkoon $f : (E, \mathscr{E}) \to ([0,\infty], \mathscr{B}[0,\infty])$ mitallinen funktio, ja määritellään, kun $\omega \in \Omega$,

$$N(\omega, f) = \int_E f(x) N(\omega, dx) \ (\leq \infty).$$

Koska f on mitallinen, voidaan tavalliseen tapaan yksinkertaisia (indikaattori-) funktioita ja monotonisen konvergenssin lausetta käyttäen osoittaa, että $N(\omega, f)$ on satunnaismuuttuja. Lisäksi

$$\mathbb{E}N(\omega, f) = \mu(f) = \int f d\mu = \int_E f(x)\mu(dx);$$

ks. tarkemmin [3, s. 127].

D.1 Poisson-satunnaismitta

Edellä nähtiin, että pisteprosessit ovat laskurimuuttujien kokoelmia. Seuraavaksi esitettävä Poisson-satunnaismitta (Poisson random measure) esiintyy monien pisteprosessien luonnollisena rajaprosessina (ks. esim. osio 1.6). Olkoon μ Radon-mitta $\mathscr{E}$:llä. Muistetaan Poisson-satunnaismitan määritelmä osiosta 1.6.1.3:

Määritelmä D.6 (Poisson-satunnaismitta (PRM))
Pisteprosessia N kutsutaan Poisson-pisteprosessiksi *tai* Poisson-satunnaismitaksi *keskiarvomitalla μ (merkitään lyhyesti PRM(μ)), jos seuraavat kaksi ehtoa täyttyvät:*

(a) Mille tahansa $A \in \mathscr{E}$ ja ei-negatiiviselle kokonaisluvulle $k \geq 0$,

$$\mathbb{P}(N(A) = k) = \begin{cases} e^{-\mu(A)} \frac{(\mu(A))^k}{k!}, & \text{jos} \quad \mu(A) < \infty, \\ 0, & \text{jos} \quad \mu(A) = \infty. \end{cases}$$

(b) Mille tahansa $k \geq 1$, jos $A_1, \ldots, A_k$ ovat keskenään erillisiä joukkoja $\mathscr{E}$:ssä, niin $N(\cdot, A_1), \ldots, N(\cdot, A_k)$ ovat riippumattomia satunnaismuuttujia.

Kohdasta (a) seuraa, että mikäli $\mu(A) = \infty$, niin $N(A) = \infty$ m.v. Lisäksi μ on N:n intensiteetti.

Propositio D.7 (Poisson-satunnaismitan olemassaolo)
Poisson-satunnaismitta PRM(μ) on olemassa.

Todistus. Todistetaan väite konstruoimalla ehdot täyttävä pisteprosessi. Seuraava konstruktio noudattaa lähdettä [3]. Tarkastellaan ensin tapausta, jossa annettu mitta μ on äärellinen, $\mu(E) < \infty$, jolloin voidaan kirjoittaa $\mu = \lambda\nu$, missä ν on todennäköisyysmitta. Olkoon

$$K, X_1, X_2, \ldots$$

riippumattomia satunnaismuuttujia määriteltynä samalla todennäköisyyskentällä $(\Omega, \mathscr{F}, \mathbb{P})$, missä K on Poisson-jakautunut satunnaismuuttuja parametrilla

$\lambda > 0$, $K \sim \text{Poi}(\lambda)$, ja $(X_i)_{i\geq 1}$ ovat iid satunnaismuuttujia tila-avaruudessa E jakaumalla ν: $\mathbb{P}(X_i \in A) = \nu(A)$, $A \in \mathscr{E}$. (Voidaan ottaa $\Omega = \mathbb{N}_0 \times E \times E \times \ldots$, missä $\mathbb{N}_0$ tarkoittaa ei-negatiivisia kokonaislukuja, varustettuna sopivalla tulo-σ-algebralla ja -mitalla.) Määritellään N ehdosta

$$N = \sum_{i=1}^{K} \delta_{X_i}, \quad K > 0,$$

ja $N = 0$, kun $K = 0$. Toisin sanoen, kun $A \in \mathscr{E}$,

$$N(A) = \sum_{i=1}^{K} \mathbb{1}_{\{X_i \in A\}},$$

kun $K > 0$, ja nolla muuten. Osoitetaan ensin, että N on pisteprosessi. Tätä varten riittää näyttää, että $N(A)$ on satunnaismuuttuja kun $A \in \mathscr{E}$. Kun $n \geq 1$,

$$\{N(A) = n\} = \sum_{k=n}^{\infty} \left(\left\{ \sum_{i=1}^{k} \mathbb{1}_{\{X_i \in A\}} = n \right\} \cap \{K = k\} \right),$$

ja siis $\{N(A) = n\}$ on mitallinen. Tapaus $n = 0$ osoitetaan samalla tavalla. Edelleen nähdään, että $N(A)$ on itse asiassa Poisson-jakautunut satunnaismuuttuja: Kun $n \geq 1$,

$$\mathbb{P}(N(A) = n) = \sum_{k=n}^{\infty} \mathbb{P}\left(\sum_{i=1}^{k} \mathbb{1}_{\{X_i \in A\}} = n \right) \mathbb{P}\left(K = k\right).$$

Nyt $B_k := \sum_{i=1}^{k} \mathbb{1}_{\{X_i \in A\}}$ on binomijakautunut parametrilla (k, p), missä $p = \mathbb{P}(X_i \in A) = \nu(A)$; B_k:n voidaan siis tulkita ilmoittavan "onnistumisten" tai osumien $\{X_i \in A\}$ lukumäärän iid elementeille X_i. Ensimmäisen termin binomijakautuneisuuden ja K:n Poisson-jakautuneisuuden nojalla saadaan

$$\begin{aligned}
\mathbb{P}(N(A) = n) &= \sum_{k=n}^{\infty} \binom{k}{n} \left(\nu(A)\right) \left(1 - \nu(A)\right)^{k-n} e^{-\lambda} \frac{\lambda^k}{k!} \\
&= \sum_{k=n}^{\infty} \frac{\left(\lambda(1 - \nu(A))\right)^{k-n}}{(k-n)!} \frac{e^{-\lambda} \left(\lambda\nu(A)\right)^n}{n!} \\
&= e^{\lambda(1-\nu(A))} \frac{e^{-\lambda} \left(\lambda\nu(A)\right)^n}{n!} \\
&= e^{-\lambda\nu(A)} \frac{\left(\lambda\nu(A)\right)^n}{n!}.
\end{aligned}$$

N on siis pisteprosessi, ja kiinteällä A $N(A)$ on Poisson-jakautunut. Tarvitsee vielä näyttää, että määritelmän D.6 riippumattomuusominaisuus (kohta (b)) pätee. Olkoon $A_0, \ldots, A_k$ avaruuden E mitallinen ositus: $A_i \in \mathscr{E}$, $A_i \cap A_j = \emptyset, i \neq j$, $\sum_{i=0}^{k} A_i = E$. Olkoon edelleen $n_i \in \mathbb{N}_0$, $i = 0, \ldots, k$ s.e. $\sum_{i=0}^{k} n_i = n$.

Kun $n \geq 1$ (tapaus $n = 0$ samaan tapaan)

$$\begin{aligned}
&\mathbb{P}(N(A_0) = n_0, \ldots, N(A_k) = n_k) \\
&\quad = \mathbb{P}(N(A_0) = n_0, \ldots, N(A_k) = n_k, K = n) \\
&\quad = \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A_0\}} = n_0, \ldots, \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A_k\}} = n_k\right) \mathbb{P}(K = n).
\end{aligned}$$

Nyt ensimmäinen todennäköisyys on multinomiaalijakautunut parametreilla (n, p), missä $p = (p_0, \ldots, p_k) = (\nu(A_0), \ldots, \nu(A_k))$, jolloin lauseke saadaan muotoon

$$= \frac{n!}{\prod_{i=0}^{k} n_i!} \prod_{i=0}^{k} (\nu(A_i))^{n_i} \frac{e^{-\lambda} \lambda^n}{n!},$$

ja huomioimalla että $1 = \nu(E) = \sum_{i=0}^{k} \nu(A_i)$ sekä $n = \sum_{i=0}^{k} n_i$, saadaan edelleen

$$= \prod_{i=0}^{k} e^{-\lambda \nu(A_i)} \frac{(\lambda \nu(A_i))^{n_i}}{n_i!} = \prod_{i=0}^{k} \mathbb{P}(N(A_i) = n_i).$$

PRM:n määritelmän (b)-kohdan osoittamiseksi olkoon nyt $A_1, \ldots, A_k$ mielivaltaisia erillisiä joukkoja $\mathscr{E}$:ssä. Asetetaan $A_0 = E - \sum_{i=1}^{k} A_i$, jolloin $\{A_i\}_{i=0,\ldots,k}$ on avaruuden E ositus. Edellisen perusteella mille tahansa ei-negatiivisille kokonaisluvuille $n_1, \ldots, n_k$ pätee

$$\begin{aligned}
&\mathbb{P}(N(A_1) = n_1, \ldots, N(A_k) = n_k) \\
&\quad = \sum_{n_0=0}^{\infty} \mathbb{P}(N(A_0) = n_0, N(A_1) = n_1, \ldots, N(A_k) = n_k) \\
&\quad = \sum_{n_0=0}^{\infty} \mathbb{P}(N(A_0) = n_0) \prod_{i=1}^{k} \mathbb{P}(N(A_i) = n_i) \\
&\quad = \prod_{i=1}^{k} \mathbb{P}(N(A_i) = n_i).
\end{aligned}$$

$N(A_1), \ldots, N(A_k)$ ovat siis riippumattomia satunnaismuuttujia, ja määritelmän ominaisuus (b) on näytetty.

Jäljellä on vielä tapauksen $\mu(E) = \infty$ käsittely. Tätä varten hajotetaan μ siten että $\mu = \sum_{i=1}^{\infty} \mu_k$ seuraavasti: Otetaan E:n ositus $\{A_i\}_{i \geq 1}$ jolle $E = \bigcup_{i=1}^{\infty} A_i$, $A_i \in \mathscr{E}$ kompakteja joukkoja, ja määritellään μ_i mitan μ rajoittumana joukkoon A_i, $\mu_i = \mu(\cdot \cap A_i)$. Nyt μ_i on äärellinen, $\mu_i(E) = \mu(E \cap A_i) = \mu(A_i) < \infty$, koska A_i on kompakti ja μ on Radon, ja edellisen kohdan perusteella voidaan rakentaa PRM(μ_i); merkitään tätä N_i:llä. Prosessit N_i, $i \geq 1$ voidaan ottaa riippumattomiksi. Määritellään $N := \sum_i N_i$ eli $N(A) = \sum_{i=1}^{\infty} N_i(A)$ kaikille $A \in \mathscr{E}$. Voidaan osoittaa, että näin määritelty N on PRM(μ) eli Poisson-pisteprosessi intensiteettimitalla μ; ks. esim. [3, s. 133-134]. □

Lisäksi PRM(μ):n todennäköisyyslain määrittävät yksikäsitteisesti määritelmän D.6 ehdot (a) ja (b); ks. [3, Proposition 3.6].

D.2 Pisteprosessien heikosta suppenemisesta

Muistetaan alaosiosta 1.6.1.4 tässä esityksessä käytetty määritelmä (ks. myös tähän liittyvä keskustelu) pisteprossien heikolle suppenemiselle:

Määritelmä D.8 (Pisteprosessien heikko suppeneminen)
Olkoon $N, N_1, N_2, \ldots$ pisteprosesseja tila-avaruudella $E \subset \bar{\mathbb{R}}^d$ varustettuna vastaavalla Borel-σ-algebralla $\mathscr{E}$. Sanotaan, että (N_n) suppenee heikosti *N:ään $M_p(E)$:ssä, $N_n \xrightarrow{d} N$, jos*

$$\mathbb{P}\left(N_n(A_1), \ldots, N_n(A_m)\right) \to \mathbb{P}\left(N(A_1), \ldots, N(A_m)\right),$$

pätee kaikilla $A_i \in \mathscr{E}$ joille $\mathbb{P}(N(\partial A_i) = 0) = 1$, $i = 1, \ldots, m$, kaikilla $m \geq 1$.

Muistetaan, että pisteprossin N sanotaan olevan yksinkertainen, jos sen jakauma keskittyy $M_p(E)$:n yksinkertaisiin pistemittoihin; tämä tarkoittaa, että $\mathbb{P}(N(\{x\}) \leq 1, \forall\, x \in E) = 1$, eli pisteiden kerrannaisuus on 0 tai 1 m.v. Joukon T osajoukkoa B kutsutaan (topologiseksi) kannaksi, kun B on sellainen kokoelma avoimia joukkoja T:ssä, että kaikki T:n avoimet joukot voidaan kirjoittaa B:n elementtien unioneina tai äärellisinä leikkauksina.

Lause D.9 (Kallenberg)
Olkoon N yksinkertainen pisteprosessi E:llä ja $\mathscr{T}$ suhteellisesti kompaktien (relatively compact) avoimien joukkojen kanta s.e. $\mathscr{T}$ on suljettu äärellisten unioneiden ja leikkausten suhteen, ja kaikilla $A \in \mathscr{T}$

$$\mathbb{P}\left(N(\partial A) = 0\right) = 1.$$

Jos (N_n), $n \geq 1$, ovat pisteprosesseja E:llä, ja kaikille $A \in \mathscr{T}$ pätee

$$\lim_{n\to\infty} \mathbb{P}\left(N_n(A) = 0\right) = \mathbb{P}\left(N(A) = 0\right) \tag{D.1}$$

ja

$$\lim_{n\to\infty} \mathbb{E}\left(N_n(A)\right) = \mathbb{E}\left(N(A)\right) < \infty, \tag{D.2}$$

niin

$$N_n \xrightarrow{d} N$$

$M_p(E)$:ssä.

Todistus. Ks. Resnick, [3, Proposition 3.22]. □

Huomautus D.10 *Kuvausta $z_N : \mathscr{E} \to [0, 1]$,*

$$z_N(A) = \mathbb{P}\left(N(A) = 0\right)$$

kutsutaan N:n nollatodennäköisyysfunktionaaliksi. Yksinkertaisten pisteprosessien N jakauma määräytyy yksikäsitteisesti, kun tiedetään $z_N(A)$ kaikilla $A \in \mathscr{T}$. Nollatodennäköisyysfunktionaalien suppeneminen, $z_{N_n}(A) \to z_N(A)$, ei kuitenkaan itsessään ole riittävä ehto pisteprosessien (N_n) suppenemiseksi heikosti yksinkertaiseen pisteprosessiin N; tarvitaan lisäoletus, joka varmistaa pisteprosessien jonon (N_n) olevan tiukka. (D.2) *on tälläinen ehto. Ks. [17, Proposition 1.22 ja Proposition 1.23] ja [3, s. 157-160].*

Lause osoittaa, että äärellisulotteisten jakaumien ja siten pisteprosessin suppeneminen voidaan näyttää yllättävän yksinkertaisella tavalla. Tyypillisesti kun tila-avaruus E on euklidinen, lauseen $\mathscr{T}$ koostuu rajoitettujen suorakulmioiden äärellisistä unioneista. Tarkastellaan seuraavaksi konkreettisuuden vuoksi ja sovelluksia silmällä pitäen tapausta, jossa tila-avaruus on väli $E = (a, b] \subset \mathbb{R}$. Seuraava lemma on suoraan lauseen D.9 erikoistapaus.

Lemma D.11 (Suppeneminen yksinkertaiseen pisteprosessiin välillä)

Olkoon (N_n) ja N pisteprosesseja tila-avaruudella $E \subset \bar{\mathbb{R}}^d$ ja olkoon N yksinkertainen. Oletetaan, että seuraavat ehdot pätevät:

$$\mathbb{E}(N_n(A)) \to \mathbb{E}(N(A)) \tag{D.3}$$

kaikilla väleillä $A = (c, d]$, missä $a < c < d \leq b$; ja

$$\mathbb{P}(N_n(B) = 0) \to \mathbb{P}(N(B) = 0) \tag{D.4}$$

kaikilla keskenään erillisten välien $(c_i, d_i]$ unioneilla $B = \cup_{i=1}^k (c_i, d_i]$, missä $a < c_1 < d_1 < \cdots < c_k < d_k \leq b$, ja kaikilla $k \geq 1$. Tällöin $N_n \xrightarrow{d} N$ $M_p(E)$:ssä.

Tarkastellaan seuraavaksi tietyissä sovelluksissa tärkeää pisteprosessien luokkaa. Olkoon

$$N_n = \sum_{i=1}^{\infty} \delta_{(\frac{i}{n}, X_{n,i})}, \quad i = 1, 2, \ldots, \tag{D.5}$$

missä $(X_{n,i})$ ovat avaruuden E iid satunnaiselementtejä. Tässä $n^{-1}i$ voidaan tulkita skaalatuksi (deterministiseksi) aikakoordinaatiksi, ja $X_{n,i}$ skaalatuksi (satunnaiseksi) paikkakoordinaatiksi.

Lause D.12 (Heikko suppeneminen Poisson-satunnaismittaan)
Olkoon $\{X_{n,i}, i \geq 1\}$ iid satunnaiselementtejä avaruudessa $(E, \mathscr{E})$, ja μ Radon-mitta $(E, \mathscr{E})$:llä. Olkoon edelleen (N_n) jono pisteprosesseja (D.5) *ja oletetaan, että N on* PRM *tila-avaruudella $[0, \infty) \times E$ keskiarvomitalla $|\cdot| \times \mu$, missä $|\cdot|$ on Lebesgue-mitta $[0, \infty)$:llä. Tällöin*

$$N_n \xrightarrow{d} N, \quad n \to \infty,$$

$M_p([0, \infty) \times E)$:ssä jos ja vain jos relaatio

$$n\mathbb{P}\left(X_{n,1} \in \cdot\right) \xrightarrow{d} \mu, \quad n \to \infty, \tag{D.6}$$

pätee E:llä.

Todistus. Ks. Resnick, [3, Proposition 3.21]. □

Huomautus D.13 *Sovelluksissa yleensä tila-avaruus on $E \subset \bar{\mathbb{R}}^d$. Tyypillisesti $E = (0, \infty]$, $E = [-\infty, \infty]$ tai $E = [-\infty, \infty] \backslash \{0\}$. Tällöin $\mu_n \xrightarrow{d} \mu$ vastaa sen osoittamista, että $\mu_n((a, b]) \to \mu((a, b])$ kaikilla $a < b$, missä $b \leq \infty$ (jos E ei sisällä nollaa, nolla ei saa sisältyä väleihin $(a, b]$).*

D.2.1 Ylitysten pisteprossin heikko suppeneminen

Osoitetaan, että osiossa 1.6.1.5 käsitelty ylitysten pisteprosessien jono suppenee heikosti Poisson-pisteprosessiin tila-avaruudella $E = (0, 1]$. Seuraava on propositio 1.59 tässä toistettuna:

Propositio D.14 (Ylitteiden pisteprosessin heikko suppeneminen) *Olkoon (X_n) jono iid satunnaismuuttujia jakaumalla F, ja olkoon (u_n) jono kynnysarvoja, s.e.*

$$n\bar{F}(u_n) = \mathbb{E}\sum_{i=1}^{n} \mathbb{1}_{\{X_i > u_n\}} \to \tau. \tag{D.7}$$

pätee jollakin $\tau \in (0, \infty)$. Tällöin ylitteiden pisteprosessien

$$N_n(\cdot) = \sum_{i=1}^{n} \delta_{\frac{i}{n}}(\cdot)\mathbb{1}_{\{X_i > u_n\}} \tag{D.8}$$

jono (N_n) suppenee $M_p(E)$:ssä heikosti pisteprosessiin N,

$$N_n \xrightarrow{d} N, \quad n \to \infty,$$

missä N on homogeeninen Poisson-prosessi tila-avaruudella E intensiteetillä τ; ts. N on $\mathrm{PRM}(\tau\,|\cdot|)$*, missä $|\cdot|$ tarkoittaa Lebesgue-mittaa E:llä.*

Todistus. Seuraava noudattaa lähdettä [2, s. 239–240]. Voidaan olettaa, että prosessi N sisältyy homogeeniseen Poisson-prosessiin $[0, \infty)$:llä. Tällöin N on yksinkertainen, ja voidaan soveltaa lemmaa D.11. Otetaan $A = (a, b] \subset (0, 1]$, jolloin

$$\begin{aligned} N_n(A) &= \sum_{i=1}^{n} \delta_{\frac{i}{n}}(A)\mathbb{1}_{\{X_i > u_n\}} \\ &= \sum_{a < \frac{i}{n} \le b} \mathbb{1}_{\{X_i > u_n\}} \\ &= \sum_{i=\lfloor na \rfloor + 1}^{\lfloor nb \rfloor} \mathbb{1}_{\{X_i > u_n\}}, \end{aligned}$$

missä $\lfloor \cdot \rfloor$ tarkoittaa kokonaislukuosaa. Havaitaan, että satunnaismuuttuja $N_n(A)$ on binomijakautunut parametreilla $(\lfloor nb \rfloor - \lfloor na \rfloor, \bar{F}(u_n))$: täten oletuksen (D.7) mukaan

$$\mathbb{E}(N_n(A)) = (\lfloor nb \rfloor - \lfloor na \rfloor)\bar{F}(u_n) \sim (n(b-a))\left(n^{-1}\tau\right) = \mathbb{E}(N(A)),$$

mikä todistaa (D.3):n (yllä $a \sim b$ tarkoittaa $\lim_{n\to\infty} \frac{a}{b} = 1$).

Täytyy vielä osoittaa (D.4). Jälleen $N_n(A)$:n binomijakautuneisuuden perusteella, sekä käyttämällä (D.7):n ekvivalenttia muotoa $\mathbb{P}(M_n \le u_n) \to \exp\{-\tau\}$, saadaan

$$\begin{aligned} \mathbb{P}(N_n(A) = 0) &= (1 - \bar{F}(u_n))^{(\lfloor nb \rfloor - \lfloor na \rfloor)} \\ &= \exp\left\{(\lfloor nb \rfloor - \lfloor na \rfloor)\ln\left(1 - \bar{F}(u_n)\right)\right\} \\ &\to \exp\left\{-\tau(b-a)\right\}. \end{aligned} \tag{D.9}$$

Kun muistetaan joukon B määritelmä (D.4):stä ja käytetään satunnaismuuttujien (X_i) riippumattomuutta, saadaan

$$\begin{aligned}
\mathbb{P}(N_n(B)=0) &= \mathbb{P}\left(N_n((c_i,d_i])=0, \quad i=1,\ldots,k\right) \\
&= \mathbb{P}\left(\max_{\lfloor nc_i\rfloor<j\le\lfloor nd_i\rfloor} X_j \le u_n, \quad i=1,\ldots,k\right) \\
&= \prod_{i=1}^{k} \mathbb{P}\left(\max_{\lfloor nc_i\rfloor<j\le\lfloor nd_i\rfloor} X_j \le u_n, \quad i=1,\ldots,k\right) \\
&= \prod_{i=1}^{k} \mathbb{P}\left(N_n((c_i,d_i])=0\right) \\
&\to \prod_{i=1}^{k} \exp\left\{-\tau(d_i-c_i)\right\},
\end{aligned}$$

missä viimeisessä kohdassa on käytetty tulosta (D.9). Toisaalta $N(B)$ on Poisson-jakautunut, koska N on Poisson-pisteprosessi, ja

$$\mathbb{P}(N(B)=0) = \exp\left\{-\tau\,|B|\right\} = \exp\left\{-\tau\sum_{i=1}^{k}(d_i-c_i)\right\},$$

eli (D.4) on osoitettu. Proposition väite seuraa nyt lemmasta D.11. □

Kirjallisuutta

[1] Coles, S.G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, New York.

[2] Embrechts, P., Klüppelberg, C. ja Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin.

[3] Resnick, S.I. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer, New York.

[4] Leadbetter, M.R., Lindgren, G. ja Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer, Berlin.

[5] McNeil, A.J., Rüdiger, F. ja Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton ja Oxford.

[6] Jacod, J. ja Protter, P.E. (2003). *Probability Essentials*. Springer, New York.

[7] Billingsley, P. (1995). *Probability and Measure*. John Wiley & Sons.

[8] Durret, R. (2010). *Probability: Theory and Examples*. Cambridge University Press.

[9] Karatzas, I. ja Shreve, S.E. (1991). *Brownian Motion and Stochastic Calculus*. Springer, New York.

[10] Revuz, D. ja Yor, M. (1999). *Continuous Martingales and Brownian Motion*. Springer, Berlin.

[11] Gariepy, R.F. ja Ziemer, W.P. (1995). *Modern Real Analysis*. PWS Publishing Co.

[12] Seppäläinen, T. (2003). *Basics of Stochastic Analysis*. Luentomoniste, Wisconsin-Madisonin yliopisto. Madison, Wisconsin.

[13] Nyrhinen, H. (2008). *Äärimmäisten ilmiöiden teoriaa*. Luentomoniste, Helsingin yliopisto.

[14] Chavez-Demoulin, V. ja Davison, A.C. (2012). Modelling time series extremes. *REVSTAT* 10(1): 109–133.

[15] Nyrhinen, H. (2011). *Riskiteoria*. Luentomoniste, Helsingin yliopisto.

[16] Daykin, C.D., Pentikäinen, T. ja Pesonen, M. (1994). *Practical Risk Theory for Actuaries*. Chapman and Hall, Lontoo.

[17] Karr, A.F. (1986). *Point Processes and Their Statistical Inference*. Marcel Dekker, New York.

[18] Fréchet, M. (1927). Sur la loi de probabilité de lécart maximum. *Société Polonaise de Mathematique. Annales.* 6: 93–116.

[19] Fisher, R.A. ja Tippett, L.H.C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Proceedings of the Cambridge Philosophical Society* 24: 180–190. (Uudelleen painettu teoksissa Fisher, R.A. (1950), *Contributions to Mathematical Statistics*, Wiley, ja Bennett, J.H (toim.) (1972), Collected Papers of R.A. Fisher, Volume II, 1925-31, The University of Adelaide.)

[20] Gnedenko, B.V. (1943). Sur la distribution limité du terme d'une série aléatoire. *Annals of Mathematics* 44: 423–453.

[21] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics* 3: 119–131.

[22] Balkema, A.A. ja de Haan, L. (1974). Residual life time at great age. *Annals of Probability* 2: 792–804.

[23] Smith, R.L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science* 4: 367–394.

[24] Bingham, N.H. (2007). Regular Variation and Probability: The Early Years. *Journal of Computational and Applied Mathematics* 200: 357–363 (J. L. Teugels Festschrift).

[25] de Haan, L. (1970). On regular variation and its application to the weak convergence of sample extremes. *Mathematical Centre Tracts* 32, Mathematisch Centrum Amsterdam.

[26] de Haan, L. (1990). Fighting the arch-enemy with mathematics. *Statistica Neerlandica* 44: 45–68.

[27] Helsingin kaupungin tulvastrategia, *Helsingin kaupunkisuunnitteluviraston yleissuunnitteluosaston selvityksiä* 2010:1.

[28] Casella, G. ja Berger, R.L. (2002). *Statistical Inference*. Duxbury, Pacific Grove, CA.

[29] Smith, R.L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* 72: 67–92.

[30] Daley, D.J. ja Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*, 2. painos. Springer.

[31] Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* 83: 9–27.

[32] Ollila, M. (toim.) (2002). Ylimmät vedenkorkeudet ja sortumariskit ranta-alueille rakennettaessa: Suositus alimmista rakentamiskorkeuksista. *Ympäristöopas* 52. Suomen Ympäristökeskus, ympäristöministeriö ja Maa- ja metsätalousministeriö.

[33] Shiryaev, A.N. (1999). *Essentials of Stochastic Finance: Facts, Models, Theory*. World Scientific, Singapore.

[34] Chavez-Demoulin, V., Davison, A.C. ja McNeil, A.J. (2005). A point process approach to Value-at-Risk estimation. *Quantitative Finance* 5(2): 227–234.

[35] Hawkes, A.G. (1971). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 33: 438–443.

[36] McNeil, A.J. ja Frey, R. (2000). Estimation of tail-related risk measures for heteroskedastic financial time series: an extreme value approach. *Journal of Empirical Finance* 7: 271–300.

[37] Artzner, P., Delbaen, F., Eber, J.M. ja Heath, D. (1999). Coherent measures of risk. *Mathematical Finance* 9: 203–228.

[38] Acerbi, C. ja Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking and Finance* 26: 1487–1503.

[39] Gourieroux, C. (1997). *ARCH-Models and Financial Applications.* Springer.

[40] Leppisaari, M. (2009). *Estimation of risk measures for heavy-tailed and heteroskedastic financial data using EVT*. Esitys 9.12.2009, Teknillinen korkeakoulu, Matematiikan ja systeemianalyysin laitos.

[41] Ljung, G.M. ja Box, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika* 65: 297–303.

[42] Daníelsson, J. ja de Vries, C.G. (1997). Value-at-Risk and extreme returns. Teoksessa *Extremes and Integrated Risk Management* (toim. P. Embrechts), s. 85–106. Risk Waters Group, London.

[43] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6): 716–723.

[44] Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics* 6(2): 461–464.